The world is random

一部真正有趣的
概率统计入门读物

我宁要模糊的正确，也不要精确的错误。
——沃伦·巴菲特

大数据时代的
概率统计学

世界是随机的

李 帅◎著

薛定谔的猫 与 庞加莱的称
科比的强悍 与 马刺的稳定
庄家的秘密 与 赌神的绝学
东野的推理 与 谷歌的预测



清华大学出版社

世界是随机的

——大数据时代的概率统计学

李　帅　著

清 华 大 学 出 版 社

北　京

内 容 简 介

这是一本写给初学者的书，目的是帮助读者理解大数据下概率统计等概念的意义，写作中以案例作先导，引起读者的兴趣和思考，在解答问题的过程中讲述知识。

本书共有9章，第1章和第2章介绍概率和随机变量的基础知识；第3章和第4章介绍统计和分布的基础知识；第5章是专门介绍赌博中的概率统计的一章，前四章的知识在这里得到了应用；第6、7、8章分别介绍了概率统计的三个重要方法——假设检验、贝叶斯定理和线性回归；第9章是漫谈概率统计。本书努力避开说教式的言辞，把知识融入故事中，在讲解知识的同时，带给读者阅读的乐趣。是一本难得的适合所有对概率统计感兴趣或者学习有需求的读者阅读。希望本书可以帮助读者更快速、更深刻地理解和应用大数据。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目(CIP)数据

世界是随机的：大数据时代的概率统计学/李帅著．—北京：清华大学出版社，2017
ISBN 978-7-302-46109-8

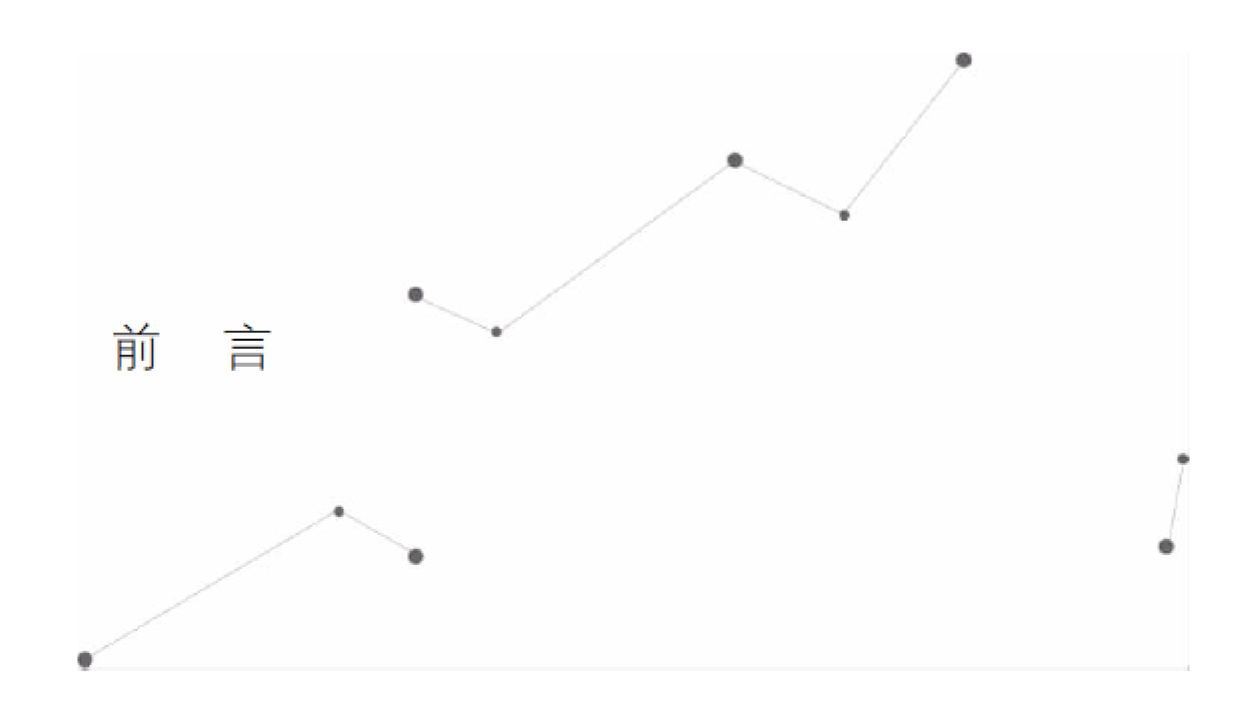
Ⅰ．①世…　Ⅱ．①李…　Ⅲ．①概率统计　Ⅳ．①O211

中国版本图书馆CIP数据核字(2016)第313733号

责任编辑：刘志彬
封面设计：汉风唐韵
责任校对：宋玉莲
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　**邮　　编：**100084
社 总 机：010-62770175　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62770175转4506
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　**印　　张：**13.25　**字　　数：**214千字
版　　次：2017年3月第1版　**印　　次：**2017年3月第1次印刷
印　　数：1～4000
定　　价：39.00元

产品编号：072671-01



前言

凯文·凯利在《失控》中曾提道，当高度互联的低级群体的数量大到一定程度时，群体特征便会涌现出来，这特征是群体中的任何个体都不具备的。比如，大量水滴汇集成河水、海水，便会产生让水滴“感到陌生”的新特征——漩涡和波浪。

2013年8月，谷歌公司提出了一个票房预测模型，该模型仅以单词搜索量为依据，便可以提前一个月预测电影的首周票房，准确度高达94%。更令人惊讶的是，这是一个简单的线性回归模型。谷歌是如何做到的呢？

人类对数据的处理已经进入大数据时代。可是，绝大多数的人，对数据统计等基本常识还在算术常识时代。这是一个科技的时代，相对于十年前和二十年前，全球市值最大最受人尊敬的公司 Top 10，全部变成了苹果、微软、Google……这些高科技公司，任何普通人都用智能手机，任何人都在享受高科技技术带来的便利。为了更好地工作和生活，我们要了解一下这些高科技技术的常识。笔者在这方面有一些经验，所以特地编写了本书，希望以比较科普和有趣的笔调，让你了解一门新的科学，甚至进入一个新的领域。

大学本科时，我曾上过“概率论”和“数理统计”两门课，

虽然完整地学习了概率统计，却只是一知半解。攻读硕士时，我在科研工作中需要用到概率统计，方才无奈地发现，当年所学已完完全全地还给了老师。我只能匆忙地自学了概率统计，勉强能应付科研工作，但心中对概率统计的很多概念仍旧一头雾水。后来，我有幸与我的妻子走到了一起，她大学本科和硕士期间都主修“应用数学”专业，在她的帮助下，我这个概率统计的门外汉终于入门了。

硕士毕业前，我和妻子共同翻译了一部英文科普读物《让你爱上数学的 50 个游戏》，这本书帮助我进一步巩固了概率统计知识，也让我萌生了写书的念头。毕业后我仍从事科研工作，参与了几个与数据分析有关的项目，发现自己对概率统计的理解仍然不够深刻。于是我一口气阅读了几本概率统计的科普书，比如《深入浅出数据分析》《深入浅出统计学》和《生活中的概率趣事》，终于搞懂了“贝叶斯定理”“假设检验”等概念。看书之余，我在“简书”上写了几篇读书心得。出版社的编辑看到我写的文章，问我是否愿意写一本概率统计的科普书，说实话，能写作一本属于自己的书是我的小小理想，既然机会来了，我怎么会拒绝呢?!

开始写作前，我为自己设定了三个原则。

一是理解而非定义。概率统计的教科书里充满了数学公式，虽然数学公式能对抽象的概念做出精确的定义，但这样的定义太晦涩，难以理解。这是一本写给初学者的书，我想帮助读者理解概念的含义，而非怎么求解某个具体问题。所以，我会用解释性的语言来描述概念，而不是给出标准的定义。这么做风险很大，但我愿意尝试，希望本书可以帮助读者更快速、更深刻地理解概念。

二是引导而非灌输。从小到大，我们都承受了太多的灌输式教育，我很庆幸，自己在灌输式教育下活了下来，但我不希望“灌输”给读者任何东西。所以，我总是以案例作先导，先引起读者的兴趣和思考，然后在解答问题的过程中讲述知识。希望这么做可以为读者减负，让读者更流畅的阅读，在轻松愉快中学到知识。

三是有趣而非无趣。很多人说，“有趣”是对一个人最高的评价。我觉得，对一本书同样如此。图书销售排行榜上，小说永远是主角，因为它们“有趣”。读者喜欢故事，不喜欢说教，这是事实，更是真理。我要努力避开说教式的言辞，把知识融入故事中，在讲解知识的同时，带给读者阅读的乐趣。

写作时，我尽量坚持这三个原则，虽然期间有过挣扎和迷茫，但最终还是完成了这本书。

本书共有9章，第1章和第2章介绍概率和随机变量的基础知识；第3章和第4章介绍统计和分布的基础知识；第5章是专门介绍赌博中的概率统计的一章，前4章的知识在这里得到了应用；第6、7、8章分别介绍了概率统计的三个重要方法——假设检验、贝叶斯定理和线性回归；第9章是漫谈概率统计。

我的阅读建议是：第1、2章合并阅读，第3、4章合并阅读，在前4章阅读完成后，再阅读第5、6、7、8、9章，后5章各自独立，不需要按顺序阅读。

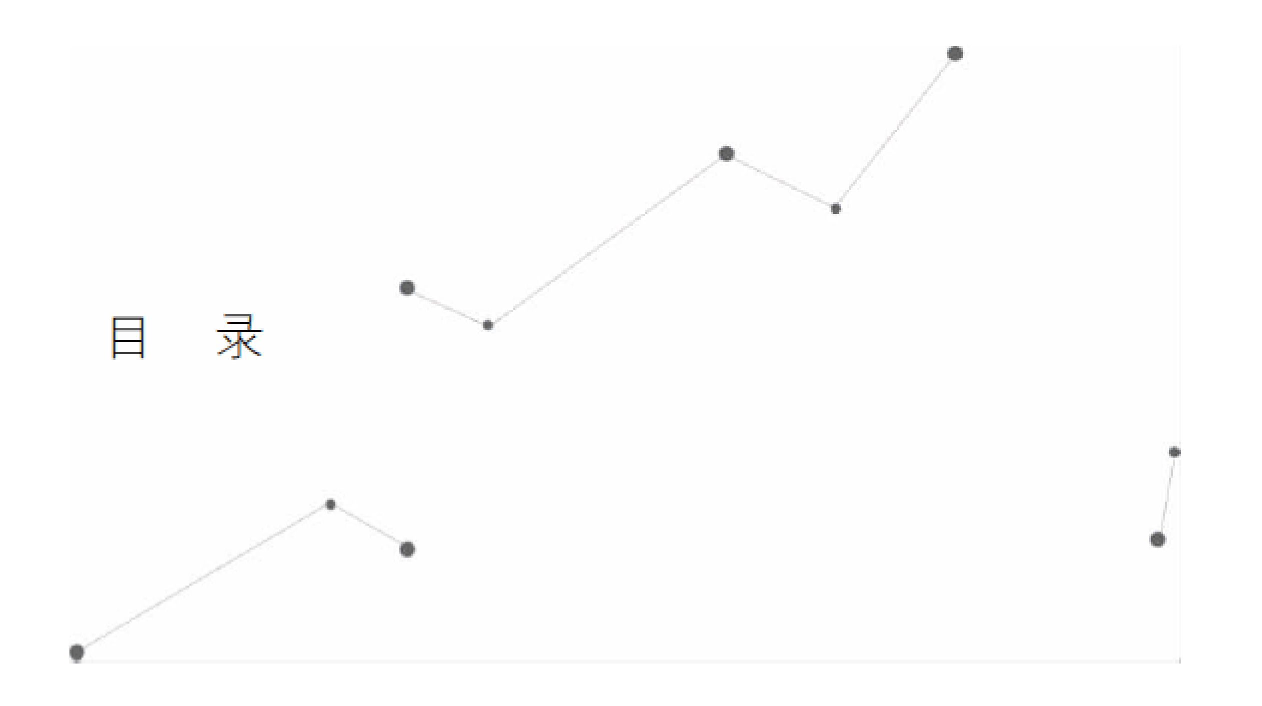
本书由李帅主笔编写，同时参与编写的还有黄维、金宝花、李阳、程斌、胡亚丽、焦帅伟、马新原、能永霞、王雅琼、于健、周洋、谢国瑞、朱珊珊、李亚杰、王小龙、张彦梅、李楠、黄丹华、夏军芳、武浩然、武晓兰、张宇微、毛春艳、张敏敏、吕梦琪等作者。在此一并感谢。

这是我的第一本书，其中难免出现错误，希望读者理解包涵，也欢迎读者批评指正。

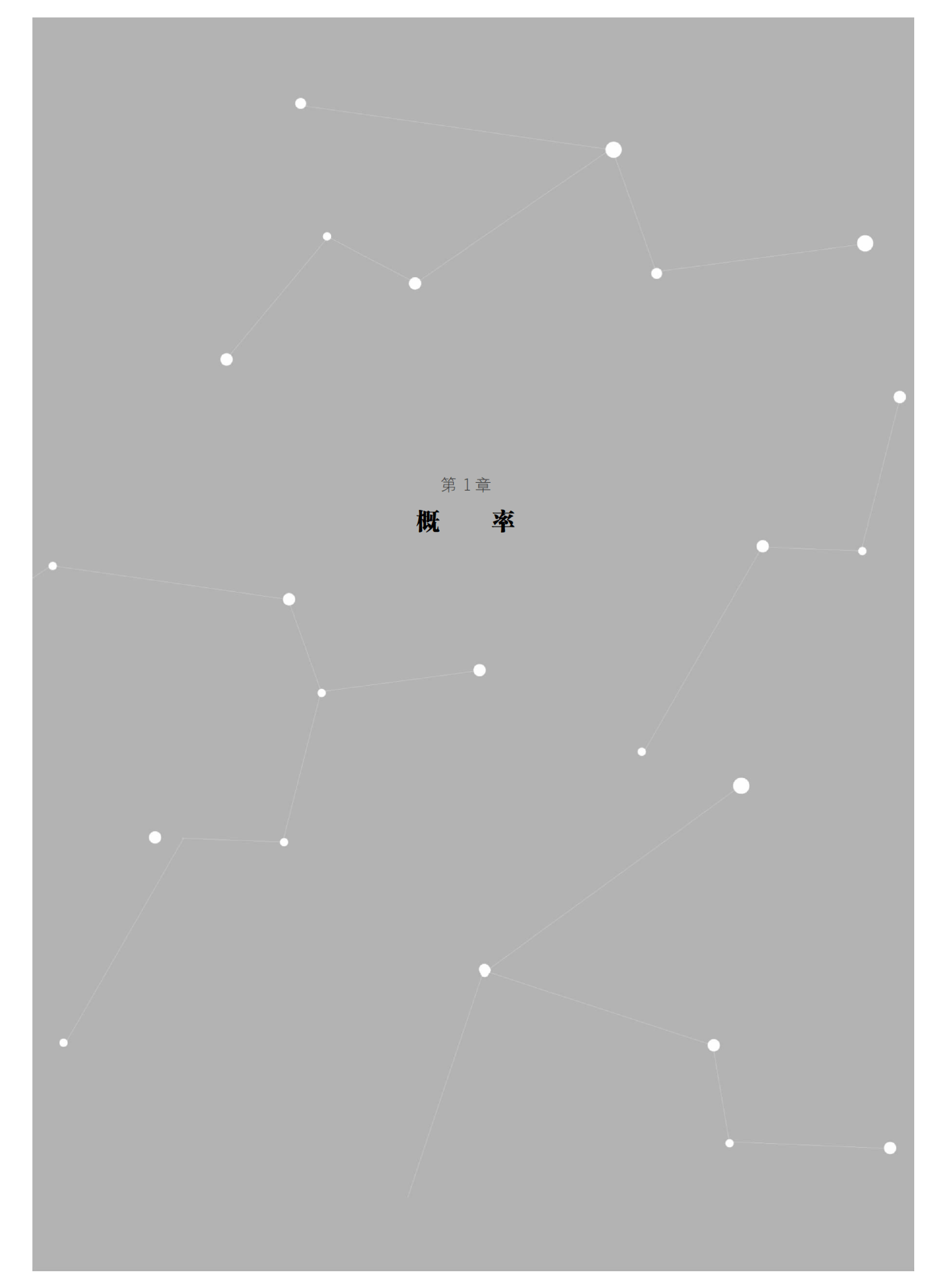
如果你读过本书，想与我沟通，欢迎通过E-mail联系我：lishuaibeijing@163com。

最后，我要感谢我的家人和朋友。感谢我的父母，陪伴我成长，帮助我养成了读书和写作的习惯。感谢我的妻子，一直理解我、陪伴我，并给我讲解了一些晦涩的数学概念。感谢刘子冲、王充山、秦培根、刘翼、孙淼、赵玮琪等老朋友，你们的支持和鼓励是我坚持写作的动力！

编　者



目　录



第 1 章

概　　率

导语：我们生活的世界，是确定的还是不确定的？自古至今，人们一直试图回答这个哲学命题。一方面我们确信，苹果熟透后会从树上掉下来；另一方面我们又无法确信，抛起的硬币落到地上时，哪一面会朝上。

1.1 生还是死：这是一个概率问题

2012 年 7 月 21 日，北京大雨倾盆，事后这一天被称为“北京 7 · 21 特大暴雨”。下午两点，我接到父亲的电话，要我赶快回东北老家。家中病危的爷爷快挺不住了。

我抓起外套出了门，冒着大雨疯狂地跑进地铁，奔向北京站。

第二天傍晚五点半，我下了火车，直奔医院。病床前，我看到瘦骨嶙峋的爷爷蜷缩在那里，已经没了意识，奄奄一息。八点整，爷爷血压骤降，医生对父亲点了点头，时辰到了。我终究没能和爷爷说上最后一句话。

后来，我常会梦到爷爷。在梦中，爷爷坐在青绿色的老式沙发上，戴着折叠

式老花镜，饶有兴致地看《城市晚报》。我似乎记得爷爷已经去世了，但又分明看到爷爷就坐在那里。那一刻，梦中的那一刻，我真的分不清爷爷是生还是死。

生死与有无、对错一样，都是鲜明对立的东西，它们看似是两条平行的直线，永不相交。然而，梦中的我却分不清爷爷是生还是死。生与死真的永无相交的可能吗？

鹰溪桥上的法克尔

下面是美国小说家安布鲁斯·布尔斯的小说《鹰溪桥上》的片段节选，故事发生在美国南北战争期间，讲述的是农场主法克尔被处以绞刑的故事。

亚拉巴马州北部的铁路桥上，一个男人站在那里，俯视着桥下二十米处那湍急的流水。这人的双手被人用绳子绑在身后，一根绳索紧紧地套在他的颈部，绳索的另一端被系在他头顶上方交叉着的架子上，一段绳子松松垮垮地垂在他的膝盖处。铁轨枕木上铺着几块木板，他和要对他行刑的一名中士和两名列兵就站在上面。

那个即将被施以绞刑的男人看起来大约 35 岁，一副平民的装扮。如果从他的举止行为来看，他像是一位庄园的农场主。他五官端正——鼻子高挺，嘴唇坚毅，额头饱满，长长的黑发顺直地披在脑后，他的眼睛大而乌黑，面目和善，人们很难想象到这人即将被施以绞刑而死。

他索性睁开了眼睛，看到了他身下的流水。"如果我能把双手挣脱，"他心里这样想着，"我就能摆脱颈上的绳索，跳到河里去，然后潜到水下躲避那些子弹，拼命地游到河岸边，钻进那里的森林，就能跑回家了。谢天谢地，我家不在他们的封锁线里，我的妻子和孩子们离他们的先头部队还有些距离。"正当这些想法在犯人脑中闪过时，上尉对中士点头示意。中士从那块木板上跨到了一边。

当法克尔从桥上径直地向下坠落时，他已经没有了意识，就像是死了一样。仿佛过了很久，颈部剧烈地挤压所带来的疼痛使他从这种状态中清醒了过来，接着就感到了窒息。他知道那条绳索已经断了，他坠入了河中，那种窒

息的感觉没有加剧。他在黑暗中睁开了眼睛，看到了他上方的一道亮光。他的两只手快速的向下拍水，使身体上浮，他感觉自己的脑袋已经浮出了水面，炫目的阳光使得他睁不开眼睛。他看到了那座桥，以及给他施以绞刑的执行者，他们正大喊着用手指向这边，子弹射到水里，离他的头只有几英寸的距离，溅起的水花打在他的脸上。

法克尔猛地向水下潜去，尽量钻到水的深处。法克尔在湍急的流水中奋力地划水，他思维清晰，四肢越发有力，心里想着："上帝保佑我，保佑我能躲过所有的子弹！"

突然，他感觉自己开始一圈圈地旋转起来，像陀螺一样。水面、河岸、树林，已经离得很远的桥，还有那军事堡垒和那些士兵，都搅到了一起，变得模糊不清。水中的一处漩涡将他卷了起来，没过一会儿，他就被水流抛到了左岸边的一堆砾石上。他喜极而泣，两手抓起泥沙，一把把的往上扬，落到自己身上，喃喃地说着一些祝福的词句。他跃身而起，迅速地往坡上的岸边跑去，钻进了那片树林。

那一天，他都依照着太阳往前走，那片树林太过茂密，像是永无尽头，他到处都找不到一个可以休息的地方，甚至都找不到一条樵夫走过的小道。夜幕降临时，他已经走得精疲力竭，可是一想到他的妻子和孩子们，他又竭力地继续向前走。最后，他终于找到了一条通往他家的路。那条路像城市里的街道那样笔直而宽阔，可却像是无人从此处通行过，路的两边没有田野，也没有房屋。他的眼睛有些肿胀，没法闭眼，口中干渴，舌头也发胀起来，他把舌头伸出口外去接触空气，感受丝丝的凉意。这条没人走过的路上全是草，这些草多么柔软，软得让他没法儿感觉到脚下的路！

他站在自己家门口，所有的一切都和他离开时一模一样。当他推开门，他看到了女人的衣裙在飘动；他的妻子还是那么的清新甜美，正从门廊中走出来迎接他。她走下了台阶，脸上带着不可言喻的笑容，那种气质简直无与伦比！啊，她是多么的美丽！他伸开双臂冲过去……

——节选自《鹰溪桥上》

读到这里，我们的心中难免会有一个疑问：法克尔究竟是死了还是逃跑了？

读到法克尔掉入水中，拼命挣扎着爬上岸时，我们相信法克尔真的逃脱了。可是，怪异的树林、无人走过的路、无法感觉脚下的路，又让人心生怀疑：难道这些是法克尔的幻觉？我们希望法克尔成功逃脱，回到家中与妻子团圆，又担心一切都是法克尔的幻觉。法克尔在我们心中仿佛是一个既可能“生”又可能“死”的人！

薛定谔的猫

要测试你是否真的了解“量子物理”，只需要问你两个问题。

第一个问题：你知道“薛定谔的猫”吗？

（我猜你会点头。）

第二个问题：你知道哥本哈根学派吗？

（别皱眉了，赶快承认不知道吧。）

大多数人都知道这只著名的猫，却不知道这只猫到底是怎么来的，没错，这只猫与哥本哈根学派有莫大的关系。

哥本哈根学派于20世纪20年代初期建立，对量子物理的创立和发展做出了很多重要贡献。学派的创始人是著名量子物理学家玻尔，主要成员包括玻恩、海森堡等知名物理学家。薛定谔也是量子物理学界的鼻祖，他提出的“薛定谔方程”为量子力学奠定了坚实的基础，至今折磨着一代又一代的大学工科男。不过，薛定谔并不是哥本哈根学派的成员，这是因为他对哥本哈根学派的理论存在质疑。为了有的放矢地提出自己的质疑，他脑洞大开地想到了一个实验——“薛定谔的猫”。

“薛定谔的猫”是一个思想实验，实验的过程是，把一只可怜的雌性小猫关在一个密室里，密室里有食物也有毒药，毒药装在瓶子里，瓶子上有一个锤子，锤子由一个电子开关控制，如果电子开关被触动，锤子就会落下，砸碎瓶子，瓶子里的有毒氰化物会毒死小猫。问题是：小猫到底是活着还是死了？

实验的关键在于，电子开关是否被触动是一个随机发生的事件，发生的概率是50%。这里的50%不是“抛硬币50%出现正面”这么简单，要产生真正的随机事件，需要使用放射性元素。在微观世界里，放射性元素的衰变是宇宙都无法预知的随机事件，一个真正的有50%概率发生的随机事件。控制电子开

关的正是放射性元素，如果放射性元素发生衰变，则开关被触动，锤子砸碎毒瓶，小猫必死。

这个问题要分两种情况讨论。

情况一：我们打开密室观察，可以确切地知道小猫是生还是死。如果放射性元素发生了衰变，那么可怜的小猫一定已经中毒身亡；如果没发生衰变，那么可爱的小猫依然活着。

情况二：我们不打开密室，由于放射性元素的衰变完全无法预测，所以小猫既可能生，也可能死，我们只能认为小猫处于“生与死”的叠加状态！

用量子物理的语言来说，当我们没有观察小猫时，小猫是被“概率云”包裹的，生与死两种状态互相叠加，形成了一个“叠加态”，当我们进入密室观察小猫时，“概率云”瞬间塌缩了，于是我们只能观察到某一种状态的小猫。

一只“既生又死”的猫？这明显违背常识。薛定谔把微观世界的叠加状态平行的移植到宏观世界中，以此质疑量子物理的“完备性”，也就是说，量子物理中的“叠加态”在宏观世界中不成立。

量子物理学家玻尔曾说：“谁要是第一次听到量子理论时没有感到困惑，那他一定没听懂。”亲爱的读者朋友，你是听懂了还是没听懂呢？

我们活在当下，感知当下，环顾四周，仿佛一切都是确定无疑的。可是，此时此刻，还有很多人、很多事是你感知不到的，对你而言，它们是“不确定的”。鹰溪桥上的法克尔和薛定谔的猫到底是生还是死？这不再是一个非此即彼的问题，在谜底揭开之前，它们既可能生，也可能死，这是一个概率问题，专门研究概率问题的学科就是——概率论。

最后，我要公布《鹰溪桥上》的结局了。

他伸开双臂冲过去，正要和那美丽的女人拥抱时，他感觉到自己的颈后遭到了重重的一击，随着一声大炮的轰鸣，他的四周亮起了炫目的白光——接着，一切都陷入了黑暗和静寂。

法克尔死了，他那折断了颈部的尸体正悬在鹰溪桥后面的横木下轻轻地摆动。

——节选自《鹰溪桥上》

1.2 随机事件：翻飞的硬币

我的家乡邻近长白山，那一年，我终于登上了长白山，见到了传说中的天池。站在山顶向下望，天池宛若一面蓝色的魔镜，静如止水，莫过如此。上山之前，很多人说，想看到天池要靠运气，没多一会儿，我就明白了此言不虚。刚刚还晴空万里、阳光普照，转瞬间就是大雾弥漫，我和父亲母亲只能手拉着手站在原地，生怕在白茫茫的雾气中走失。再过一会儿，雾气缓缓消散，正当大家拿出相机要继续拍照时，乌云袭来，风雨大作，我们纷纷披上雨衣，站在寒风中瑟瑟发抖。那是我第一次感到大自然的风云变幻。

自古至今，人们都在试图回答一个哲学命题：我们生活在一个确定的世界还是不确定的世界？我们很确信，苹果熟透了，会从树上掉下来，但我们又不能确定，抛起的硬币落到地上时，哪一面会朝上。对此，哲学领域有两种不同的论断。

决定论：它是指自然界和人类社会普遍存在着客观规律和必然的因果联系，也就是说，如果我们能够发现和理解所有的客观规律和因果联系，自然界和人类社会的任何变化都是可以预知的，我们之所以还做不到，是因为我们对客观规律的认识还不够。

非决定论：与决定论相对，非决定论否认自然界和人类社会普遍存在着客观规律和必然的因果联系，认为事物的发展变化是没有客观规律的，是由事物内在的“自由意志”决定的，也就是说，人们可以自由支配自己的行为，却无法预言客观事物的发展变化和其他人的行为。

我们似乎更容易认同非决定论，毕竟世界如此纷繁复杂，我们只能控制自己，很难预知未来。但我们不能轻易否定决定论，抛开两个论断的对错之争，决定论为我们认识世界提供了新的思路。下面，我们就来做一个“抛硬币”的思想实验。

思想实验：抛硬币

抛硬币是大家十分熟悉的小把戏，足球比赛前，裁判会用抛硬币的方式让

双方挑边，大家似乎默认抛出的硬币落到手上或地上时，正面和反面朝上的可能性是相同的。但是，决定论的支持者们对此表示怀疑，他们提出了如下的思想实验。

实验 1.0

假定有一台超高速摄像机和一台超级力学计算器，摄像机自带摇臂，可以跟拍动态画面，并对拍摄到的画面进行实时分析，分辨画面中的物体，提取物体的运动参数，这些参数又被实时的传输到力学计算器，力学计算器可以根据此前的数据计算出物体下一时刻的运动状态。

我们用超高速摄像机对准手上的硬币，然后，抛起硬币！超高速摄像机与硬币一起向上升，又一起向下降，最后，在硬币即将落到手上时，力学计算器输出了计算结果：正面向上。你展开手掌，露出了硬币，果然是正面。

我们在实验中加入了一位超级观察员——由超高速摄像机和超级力学计算器组合而成。只要你不是魔术师，也不刻意作弊，在硬币即将落到手上时，超级观察员一定可以准确地告诉你硬币的哪一面向上。请问：抛硬币的结果是随机的吗？

我的回答依然是：随机的。原因是，硬币在运动过程中，可能受到各种因素的干扰，力学计算器只能做出短时间的预测，所以，超级观察员只能在硬币即将落到手上时，才能计算出硬币哪一面向上，因此，在硬币抛起时，即使是超级观察员也无法预测硬币的哪一面向上。为了反驳这两点，我们将思想实验升级为 2.0 版。

实验 2.0

在实验 1.0 的基础上，我们加入如下条件：一是每次硬币抛掷的周围环境都一样；二是你的手升级为超级机器手，内置力学传感器，你抛起硬币时对硬币施加的力全部会被记录在传感器的芯片中，同时，超级机器手还可以自由设定抛硬币使用的力，也就是说，你可以复现曾经出现过的硬币抛掷过程。再次请问：抛硬币的结果是随机的吗？

这时，我有些语塞了，在这样的条件下，如果我们利用超级机器手重复此前的某一次抛掷，那就意味着，在硬币刚刚抛出时，我们就知道了结果，这时，抛硬币的结果是确定的！如果我们利用这套装置不断进行抛硬币练习，就会收集越来越多的硬币抛掷结果，然后，这只超级机器手就会成为一个开关，它

既可以再现过去的抛掷过程，准确预言抛掷结果，也可以进行一次新的抛掷，让结果随机出现。这只超级机器手掌控着一切，仿佛“造物主”一样！

决定论的极限表达是“造物主”，造物主知晓一切，造物主决定一切，造物主预知一切。这种宗教化的解释自然不在我们的讨论范围内，但“决定论”赋予我们一个很有价值的思想：不断探索自然，不断寻找客观规律。试想，在牛顿发现万有引力之前，已有千千万万个苹果落到了地上，难道我们该认为，这些苹果拥有“自由意志”，竟然不约而同地冲向地面吗？这个看似必然发生的事件，正是万有引力定律引起的，对这个确定性事件的解释，让我们对大自然的认识更加深刻，也正是“决定论”指引我们不断探索下去。

度量随机事件

我们从思想实验中跳脱出来，回到现实世界。在现实世界中，每时每刻都在发生各种各样的事情，有的事像苹果落地一样，有确凿无疑的结果，而有的事却像抛硬币一样，无法预知结果。数学家们既不是决定论者，也不是非决定论者，他们从数学的角度审视万事万物，概率论由此而来。

抽象地讲，概率论站在无知者和造物主之间审视世界，力图从现实世界中发现客观规律，帮助我们更深刻的认识现实世界。

在概率论的世界里，抛硬币、掷骰子等被统称为随机试验，每一个随机试验都会有一个或多个可能的结果，一个结果或某些结果的组合称为随机事件。

举例来说，抛硬币是一个随机试验，抛硬币可能的结果有两个：正面和反面。我们用一个大写字母来代表随机事件，那么，我们可以得到如下的四个随机事件。

A：抛硬币出现正面

B：抛硬币出现反面

C：抛硬币出现正面或反面

D：抛硬币既不出现正面也不出现反面

随机事件 C 和随机事件 D 往往会给初学概率论的人带来困扰，随机事件 C 根本就不是“随机”事件，分明就是一定会发生的确定性事件，随机事件 D 正相反，是一定不会发生的事件，自然也不是“随机”事件。概率论是一门完备的

科学，它要涵盖所有的事件，而不是只研究那些“随机”事件，为此，我们需要一个度量随机事件的工具——概率。

概率，用于度量随机事件发生的可能性，是个定量指标，用大写字母 P 来表示。例如，随机事件 A 发生的概率是 50%，可以写成：

$$P(A) = 50\%$$

概率有以下两个特性：

(1) 概率是非负的，即对于任意随机事件 A，$P(A) \geqslant 0$；

(2) 对于任一随机试验，我们假定所有可能的结果有 n 种($n>0$)，分别记为 $A_1, A_2, \cdots, A_n$，如果这些结果两两之间都不可能同时出现，则 $P(A_1) + P(A_2) + \cdots + P(A_n) = 1$。

事实上，在概率论所描述的数学世界中，所有的事件都是随机事件，如果一个事件不可能发生，我们认为它发生的概率是 0，如果一个事件必然发生，我们认为它发生的概率是 1。下面我们举两个有争议的例子。

随机事件 A：公鸡下蛋。

这违背常识，不可能发生，$P(A)=0$。

随机事件 B：人终有一死。

这是个客观事实，必然发生，$P(B)=1$。

就大多数人的认知，这两个概率是正确的。可是，生物学家或许会质疑这两个概率，甚至罗列一长串的生物新技术来反驳这两个概率。没错，我承认这两个概率可能是错误的，正如崔健唱的那样：“不是我不明白，这世界变化快。”世界在变化，概率也在变化，唯一不变的是：所有的事件都是随机事件。

1.3　条件概率：门后的老山羊与豪车

一个囚犯站在法官面前听候判决。法官严肃地说：“我不得不做出最严厉、最残酷的判决，这就是绞刑。这个严酷的刑罚必须执行，不可更改。除此之外，我唯一的决定权是安排你的行刑日期，对此，我一直在两个方案之间犹疑。”

“最简单、最直接的方案是判决即刻生效，马上执行，但这个判决对你太仁

慈了,你完全没有感受到惊恐害怕。因此,我现在决定:在下周 7 天中的某一天,我会在日出时安排执行绞刑。我绝不会提前告诉任何人,我会在哪一天安排绞刑,所以,我保证你不可能事先知道,自己将在哪一天被绞死。每个夜晚,你都将在担惊受怕中入睡,这是对你最大的惩罚。”

法官宣判完后,囚犯绝望了,他转过头去,居然看到他的律师露出了微笑。走出法庭后,律师对囚犯说:“他们不能绞死你了,”他解释道,“按照法官的安排,下周 7 天中的某一天,他会在日出时分执行绞刑,而且他们保证不会提前让你知道。因此,他们不能在星期六绞死你,因为星期六是一周的最后一天,如果星期五的早晨,你还没有被绞死,你就知道了行刑日期必然是星期六。这与法官的安排是矛盾的,因为他的计划是不让你知道行刑日期。”

“所以,他们最晚只能在星期五绞死你,这一点没问题吧。”囚犯对此表示赞同。“既然星期六已经排除了,星期五就成了可以绞死你的最后一天,按照同样的逻辑,如果你星期四早上还没被绞死,那么你一定会在星期五被绞死,这又与法官的安排矛盾。你明白了吗?依照同样的逻辑,我们还可以排除星期四、星期三,我们可以排除每一天!法官把自己套住了!这个判决不可能执行!”

囚犯心情愉快地度过了星期一,星期二早晨,他从美梦中醒来,然后被押赴刑场,绞死了。

这是一个经典的悖论——意外绞刑悖论,它还有很多种表现形式,比如老师突袭考试、紧急消防演习等。正如律师所言,如果法官严格的执行判决,囚犯将不会被绞死,然而,法官在公布判决结果时已经下定决心:绞刑必须执行,只有在这个前提下,才能体现出悖论的思辨色彩。哲学家迈克尔·斯克里文这样评论意外绞刑悖论:“逻辑的力量遭到事实的否决,我觉得这正是此悖论的迷人之处。可怜的逻辑学家念着过去屡试不爽的咒语,但事实上这个怪物听不懂咒语,执意前行。”

我们用概率论分析一下这个悖论。在法官说到,要在一周 7 天中的某一天处死囚犯时,囚犯在一周 7 天的任何一天被执行绞刑的概率都是 1/7,而当法官说到,囚犯不会知道绞刑在哪一天执行时,概率发生了变化,周六执行绞刑的概率原本是 1/7,此时却降为了 0,因为周六执行绞刑违背了“囚犯不知道绞刑在哪一天执行”的条件。一个前提条件,改变了事件发生的概率,这就

是——条件概率。

“三门问题”

“三门问题”是一个知名的概率问题，这个问题刚好用到了“条件概率”，我们一起来看看，条件概率是如何帮助参赛者提高获胜机会的。

蒙提霍尔是一个美国电视节目的主持人，他曾主持过一个有趣的游戏节目，叫作“Let's make a deal”。节目中有三扇关闭的大门，其中一扇门的后边是一辆豪车，另外两扇门的后边各藏着一只老山羊。如果参赛者最终选定的门的背后是豪车，参赛者可以开着豪车回家，如果是老山羊，参赛者将空手而归。节目开始后，蒙提霍尔让参赛者从三扇关闭的门中随便挑选一扇，然后，蒙提霍尔会从剩下的两扇门中打开一扇，门后定会出现一只老山羊，因为，蒙提霍尔知道豪车藏在哪扇门的后边。此时，蒙提霍尔会给参赛者一个改选的机会，如果你是参赛者，你会改选另一扇门还是坚持最初的选择？

我猜你此刻在想：蒙提霍尔知道豪车在哪，我可不知道，所以选哪扇门都一样嘛，改或者不改是一样的，非要我决定改还是不改的话，抛硬币好了。

节目中的参赛者也是这么想的，所以他们有的坚持不改，有的摇摆不定之后改选了另一扇门。这个游戏还包含另一层心理层面的因素，如果参赛者不改变自己最初的选择，即使他们没有得到豪车，也会用“坚持自我”来安慰自己，而如果他们改选另一扇门却落了个空，则会懊恼不已，因为他们把到手的豪车拱手送了出去！看起来，不改变自己最初的选择是对的。“不变初衷”“坚持自我”，多么励志的想法！

然而，科学不相信励志。下面，我就来告诉你，为什么“坚持自我”是错误的。

这个问题中的条件有些复杂，为了由浅入深的展开分析，我们对前提条件做一个简化：假设主持人不知道哪扇门后边是豪车，也就是说，在参赛者选择完一扇门后，主持人在剩下的两扇门里随机挑选一扇。此外，为了方便起见，我们把两只老山羊分别记为公山羊和母山羊，很显然，这样不会影响计算结果。

在这样的前提条件下，我们把所有可能的情况列出来，一共有 6 种可能的

情况，即 6 个随机事件，如表 1-1 所示。

表 1-1 “三门问题”的所有可能情况

随机事件	参赛者第一次选择的门	主持人选择的门	剩下的最后一道门
A	公山羊	母山羊	豪　车
B	公山羊	豪　车	母山羊
C	母山羊	公山羊	豪　车
D	母山羊	豪　车	公山羊
E	豪　车	公山羊	母山羊
F	豪　车	母山羊	公山羊

现实中，主持人并非随机选择了一扇门，他只会选择公山羊或母山羊面前的那扇门，所以，随机事件 B 和随机事件 D 不可能发生！也就是说，当参赛者第一次选择了公山羊或者母山羊时，主持人根本没有选择的余地，他必须选择另一只山羊，留下豪车，这时，参赛者应该改变初衷，选择另一扇门；当参赛者第一次选择了豪车时，主持人一定会留下一只老山羊，这时参赛者不应该改变初衷。

因此，在下面三种情况下，参赛者会获得豪车。

参赛者选择公山羊⇒主持人选择母山羊⇒参赛者改选另一扇门⇒参赛者获得豪车

参赛者选择母山羊⇒主持人选择公山羊⇒参赛者改选另一扇门⇒参赛者获得豪车

参赛者选择豪车⇒主持人选择母山羊或公山羊⇒参赛者不改变选择⇒参赛者获得豪车

这三种情况包含的一个重要信息是：只要知道了参赛者第一次选择的门后是什么，就知道了参赛者是否应该改选另一扇门。下面，我们来计算参赛者第一次选择的三种可能的结果出现的概率。

设定：

随机事件 A_1：参赛者第一次选择公山羊；

随机事件 A_2：参赛者第一次选择母山羊；

随机事件 A_3：参赛者第一次选择豪车。

我们知道，参赛者第一次的选择是完全随机的，因此：

$$P(A_1) = P(A_2) = P(A_3)$$

并且：

$$P(A_1) + P(A_2) + P(A_3) = 1$$

因此可以得到：

$$P(A_1) = P(A_2) = P(A_3) = 1/3$$

只有当随机事件 A_3 发生时，参赛者才应该坚持自己的选择，而随机事件 A_3 发生的概率只有 1/3，所以，我们得到的结论是：改选另一扇门，有 2/3 的可能得到豪车，反之，则只有 1/3 的可能得到豪车。

重新审视分析过程，我们会发现，这个游戏有趣的一点就在于：在你随机选择一扇门之后，主持人为你去掉了一个错误答案。有了这个前提条件，参赛者获胜的概率提高了，这就是“条件概率”的神奇之处！

条件概率

条件概率，是针对两个或两个以上的随机事件提出的概念，我们设定任意两个随机事件为 A、B，那么，在 A 已经发生的前提下，B 发生的概率就称为条件概率，记为 $P(B|A)$。

概率具有非负性，条件概率是概率的一个类别，因此同样具有非负性，即对于任意随机事件 A 和随机事件 B，$P(B|A)\geqslant 0$。

要研究两个随机事件之间的关系，首先要弄清楚，两个随机事件的概率之间可以进行哪些数学运算，下面我们来介绍概率的加减乘除法则。

首先，我们要定义两个概念：

和事件：随机事件 $A \cup B$ 称为 A 和 B 的和事件，它表示随机事件 A 或随机事件 B 中至少有一个发生；

积事件：随机事件 $A \cap B$ 称为 A 和 B 的积事件，它表示随机事件 A 和随机事件 B 同时发生。通常地，我们把 $A \cap B$ 简写为 AB。

其次，我们来学习概率的加法和乘法。

概率加法：对任意两个随机事件 A 和 B，有

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

概率乘法：对任意随机事件 B 和满足 $P(A) > 0$ 的随机事件 A，有

$$P(AB) = P(B \mid A) \cdot P(A)$$

概率的加法和乘法就是概率论中的四则运算，很多概率问题的计算都需要使用这两种运算，本书后边的内容也会反复使用它们。这里需要说明的是，概率加法和乘法的证明不在本书的讨论范围内，我们把它们当作数学中的四则运算一样使用就可以了。

细心的读者会发现，概率乘法中出现了条件概率 $P(B|A)$，事实上，概率乘法的另一种表达方式正是条件概率的数学定义。

设随机事件 A 和 B，满足 $P(A) > 0$，则

$$P(B \mid A) = P(AB) / P(A)$$

定义为随机事件 A 发生的前提下随机事件 B 发生的条件概率。

关于条件概率，我们要讨论的最后一个问题是：对于某个随机事件 B 和任意随机事件 A，$P(B|A)$和 $P(B)$之间的大小关系是怎样的？

这个问题会让人在一瞬间产生两种截然相反的想法。有些人在想：已知条件越多，事情发生的概率应该越大，所以 $P(B|A) \geqslant P(B)$；另一些人在想：已知条件越多，对事件发生的限制也越多，事件发生的概率也越小，所以 $P(B|A) \leqslant P(B)$。

我们用一个生活中的事例来解释一下。

2015 年，北京空气质量达标 186 天，占全年的 51%。假定今天是 2016 年 1 月 1 日，如果我让你预测一下 3 月 1 日会是雾霾天还是晴天，你会怎么回答？谁都不可能提前两个月预测一天的雾霾情况，所以，你只能回答 51%，看起来，这跟抛硬币没什么区别。

时光如箭，转眼到了 2 月 29 日。假如，今天白天狂风大作，夜幕降临时，风停了，你一定会感到欣慰：“明天肯定是好天气！”又如，今天雾霾压城，直到夜里仍不见好转，你会在睡前默默地给全家人准备好口罩，你知道，明天肯定还是雾霾天。

我们用概率语言重新描述上面的事例。

随机事件 B：2016 年 3 月 1 日，北京城是个雾霾天。

随机事件 A_1：2016 年 2 月 29 日，北京白天刮大风，晚上风停了。

随机事件 A_2：2016 年 2 月 29 日，北京全天雾霾严重。

根据常识，我们得到了下面的结论：

$$P(B \mid A_1) < P(B)$$

$$P(B \mid A_2) > P(B)$$

所以说，$P(B|A)$和 $P(B)$没有确定不变的大小关系，前提条件对随机事件产生的影响无法预测！

1.4　独立事件：反复抛起的硬币

有这样一个谜题：小明一家四口正在沙滩上享受假期，这时，小明和妹妹为了一个美丽的贝壳争执起来，他们俩都想得到贝壳，谁也不让谁，只好找来父亲。父亲没法说服这对兄妹，只能用一种“公平的方式”来决定贝壳归谁——抛硬币。可是，父亲手上没有硬币，只有几个汽水瓶瓶盖，父亲想要用瓶盖代替硬币，可是，抛瓶盖出现正面和反面的概率未必相同。请问，父亲该怎么办呢？

或许你已经知道答案了，如果你还没想明白，先把这个小谜题放在一边，我们一起来学习概率论中的一个很独特的概念——独立事件。

独立事件的含义

通俗地讲，彼此没有任何关联的事件称为独立事件。比如，你和我各自抛一枚硬币，你抛硬币出现正面和我抛硬币出现正面是两个毫不相干的随机事件，此时，我们就称这两个事件彼此独立，互为独立事件。

独立事件看起来很容易理解，实际上，人们常常搞不清楚它的含义。下面，我们就来讨论一下独立事件真正的含义。

某日，一架小型客机在靠近机场的居民区坠落，所幸没有造成人员伤亡。记者们第一时间赶到事故现场，采访了机场总经理。为了安抚大家的情绪，也为了保全机场的声誉，机场总经理这样说道：“从统计学上讲，人们应该感到更放心，因为再次发生类似事故的可能性相比此前大大减小了。”

毫无疑问，这段采访应该入选“史上最差危机公关”的榜单。历史上有很多血淋淋的事件都可以反驳这种愚蠢至极的说法。纽约时间 2001 年 9 月 11

日早8时37分,美国航空公司11次航班被劫持,8时46分,这架波音767飞机以490千米/小时的速度撞向世贸中心北楼。要知道,在此之前,美国发生飞机撞楼事件的概率仅为0.005%,如果按照那位机场总经理的说法,世贸中心第一次被撞之后,几乎不可能再发生类似事件了。而事实是,恐怖分子随后驾驶另外两架波音飞机撞击了世贸中心南楼和五角大楼。除此以外,我们还能列举出很多"祸不单行"的事实。面对恐怖袭击或者意外事故,我们要做的不是拿概率理论来蒙骗大众,而是应该找出事故原因,避免类似的惨剧再次发生。

那么,这位机场总经理到底错在哪儿呢?他混淆了两件事:一个随机事件发生两次和一个随机事件再次发生。以飞机失事为例,设

随机事件A:该机场飞机失事

根据该机场的营运历史,$P(A)=0.01\%$

我们假设两次不同的事故之间是互相独立的,那么,该机场发生两次飞机失事事故的概率是:

$$P(A)\cdot P(A)=0.000\,001\%$$

这个概率的确远低于$P(A)$。可是,在飞机失事已经发生的时候,飞机再次失事的概率依然是:

$$P(A)=0.01\%$$

因为事故之间是彼此独立的。如果两者彼此存在关联,这个概率甚至会变得更大。

对独立事件还有另一种常见的误解。

请你快速回答:抛硬币时,"出现正面"和"出现反面"互相独立吗?

我希望听到肯定的回答,这样我就可以纠正你的错误了!关于独立事件的第二个误解就是:把不能同时发生的事件当作互相独立的事件。"正面"和"反面"的确不可能同时出现,它们看起来互不侵犯,难道不是互相独立吗?答案是否定的。因为独立事件的含义是,当一个随机事件发生时,不影响另一个随机事件发生的概率。如果抛硬币出现了正面,那么,出现反面的概率会从50%降为0!

关于独立事件,我们需要记住以下三点:

(1)一个随机事件发生两次的概率不等于一个随机事件再次发生的

概率；

（2）不可能同时发生的事件不是互相独立的；

（3）独立事件的含义是，不论一个随机事件发生还是不发生时，都不会影响另一个随机事件发生的概率。

独立事件的数学表达

还记得概率乘法吗？

$$P(AB) = P(B \mid A) \cdot P(A)$$

我们刚刚学到，独立事件的含义是，当一个随机事件发生时，不影响另一个随机事件发生的概率。这听起来很像条件概率的定义，实际上，这句话等价于下面的数学表达式：

$$P(B \mid A) = P(B)$$

将这两个表达式合并起来，就可以得到，当随机事件 A 和随机事件 B 互相独立时，

$$P(AB) = P(B) \cdot P(A)$$

上面的表述前后颠倒一下，就是独立事件的定义。

设 A 和 B 是两个随机事件，如果满足

$$P(AB) = P(B) \cdot P(A)$$

则称 A 和 B 互相独立，或称 A 和 B 互为独立事件。

这是两个事件相互独立的定义，那如果是三个事件呢？

设 A、B、C 是三个随机事件，如果满足如下等式：

$$P(AB) = P(B) \cdot P(A)$$

$$P(AC) = P(A) \cdot P(C)$$

$$P(BC) = P(B) \cdot P(C)$$

$$P(ABC) = P(A) \cdot P(B) \cdot P(C)$$

则称 A、B、C 互相独立。

由此可以推论出 n 个事件互相独立的定义，请读者们自行脑补。

本节的最后，我要告诉你那个小谜题的一个参考答案：扔两次瓶盖，出现“正面、反面”，贝壳归小明；出现“反面、正面”，贝壳归妹妹；出现其他情况，父

亲重新扔，直到贝壳有了归属为止。因为每次扔瓶盖是互相独立的，所以，出现“正面、反面”和“反面、正面”的概率一定是相等的，独立事件帮助我们实现了公平。

1.5 全概率法则：英超冠军争夺战

现代足球的百年历史画卷上留下过很多“草根逆袭”的神话，“70 后”会追忆1992 年欧洲杯的“丹麦童话”，“80 后”依稀记得 2004 年欧洲杯的“希腊神话”，我倒觉得，像欧洲杯这样的淘汰赛具有很大的偶然性。真正有实力的黑马当属 1997—1998 赛季的凯泽斯劳滕队，他们在升级到甲级联赛的第一个赛季就力压德甲霸主拜仁慕尼黑，获得了联赛冠军，在当时被认为“难后有来者”。然而，总有人要挑战不可能，来自英超联赛的小球会莱斯特城队很可能重演草根逆袭的神话。

莱斯特城队的逆袭

2015—2016 赛季的英超联赛，可谓翻天覆地，有一句话能最贴切的描述英超的现状——本想保级的队伍现在在争冠，本想争冠的队伍现在在保级，只有阿森纳实现了“争四”的目标。我们就来一起聊聊那支“本想保级，却在争冠”的队伍——莱斯特城队。

莱斯特是英格兰中部的一座城市，位于伦敦西北 156 公里，人口约 32 万。莱斯特城足球俱乐部成立于 1884 年，绰号“狐狸”，他们于 1890 年加入英格兰足球协会，在中部地区联赛里混迹了三年后，他们于 1894 年夺得亚军，获得了参加全国乙级联赛的资格。1908 年，莱斯特城队获得乙级联赛亚军，终于升入甲级联赛，然而，由于实力不济，他们很快便降级了。此后的多年，他们一直在甲级联赛和乙级联赛中徘徊，成绩不温不火。上赛季，他们从英甲联赛升入英超联赛，勉强完成了保级任务，留在了英超。夏季休赛期，时任莱斯特城主帅的皮尔逊因为儿子在泰国曝出性丑闻，被球队的泰国老板愤而解职。随后，他们请来了老帅拉涅利。

拉涅利，意大利人，绰号“补锅匠”，执教履历丰富，执教风格保守。“当我和球队谈话时，我发现他们害怕意大利战术，他们看起来不怎么相信我，我自己也是。”在近日的采访中，拉涅利谈起了刚接手莱斯特城队时的情景，“我认为一个教练最重要的是围绕自己球员的特点构建球队，所以我对球员们说，我信任你们，我不会多说战术的事。英国的比赛强度超高，几乎能把球员榨干，他们需要时间恢复。我要确保球员们每周有两天完全与足球无关，这是我在第一天就对他们强调的，这是一种信任。”

正如拉涅利所说，他对球队充分信任。拉涅利上任后，基本保留了球队的原班人马，包括助理教练团队。这使得球队很快度过了磨合期，球员们也踢的更自信。正是主帅的信任和球员的自信让莱斯特城队踢出了十分高效的足球，他们一路过关斩将，踢的霸气十足。曾经的英超五强中，只有阿森纳队在勉强追赶莱斯特城队，然而，圣诞节过后，“争四魔咒”再度降临，温格的球队无可挽回的滑向第 4 名，莱斯特城队却依旧坚挺。

“英超赛季快过半了，占据积分榜头名的是一支叫莱斯特城的球队。一年前的圣诞节，他们排名垫底，濒临降级。”面对媒体的赞扬、调侃或者质疑，莱斯特城队的教练和队员始终在强调：“我们的目标是取得 40 个以上的积分，确保保级成功。”

莱斯特城队的低调务实不是没有理由的，从勉强保级，到争夺冠军，更何况是在竞争激烈的英超联赛，这实在是天方夜谭。然而，在刚结束的英超第 31 轮，莱斯特城队 1∶0 小胜水晶宫队，将自己的领先优势保持在 5 分，随着联赛轮数逐渐减少，这样的优势促使莱斯特城队的夺冠概率变得越来越大。比赛中，莱斯特城队球迷已经在看台上高唱起“我们将要赢得英超冠军”的口号。其实，无论莱斯特城队能否最终夺冠，我们都在内心深处成了莱斯特城队的球迷，正如老帅拉涅利所说：“莱斯特能夺冠吗？我不知道，但能被问到这个问题就足够美妙了。在这个金钱衡量一切的时代，我们给了每个人希望。”

莱斯特城队的夺冠概率

假如我们都是莱斯特城队的球迷，我们一定特别想知道，莱斯特城队夺冠的概率到底有多少。表 1-2 是英超联赛截至第 31 轮的积分榜，表 1-3 是莱斯

特城队未来赛程。

表 1-2 英超联赛 2015—2016 赛季积分榜(截至第 31 轮)

排名	球　队	场次	积分	胜	平	负	进球	失球	净胜球
1	莱斯特城	31	66	19	9	3	54	31	23
2	热刺	31	61	17	10	4	56	24	32
3	阿森纳	30	55	16	7	7	48	30	18
4	曼城	30	51	15	6	9	52	32	20
5	西汉姆联	30	50	13	11	6	47	35	12
6	曼联	30	50	14	8	8	38	27	11
7	南安普敦	31	47	13	8	10	41	32	9
8	斯托克城	31	46	13	7	11	34	37	－3
9	利物浦	29	44	12	8	9	45	40	5
10	切尔西	30	41	10	11	9	45	41	4
11	西布朗	30	39	10	9	11	30	37	－7
12	埃弗顿	29	38	9	11	9	51	41	10
13	伯恩茅斯	31	38	10	8	13	38	50	－12
14	沃特福德	30	37	10	7	13	30	32	－2
15	斯旺西	31	36	9	9	13	31	40	－9
16	水晶宫	30	33	9	6	15	32	40	－8
17	诺维奇	31	28	7	7	17	32	54	－22
18	桑德兰	30	26	6	8	16	36	55	－19
19	纽卡斯尔	30	25	6	7	17	29	55	－26
20	阿斯顿维拉	31	16	3	7	21	22	58	－36

表 1-3 莱斯特城队未来赛程

轮次	主　场	客　场
第 32 轮	莱斯特城	南安普顿
第 33 轮	桑德兰	莱斯特城
第 34 轮	莱斯特城	西汉姆联
第 35 轮	莱斯特城	斯旺西
第 36 轮	曼联	莱斯特城
第 37 轮	莱斯特城	埃弗顿
第 38 轮	切尔西	莱斯特城

莱斯特城队能否夺冠不仅与自身的比赛结果有关。还与其他球队的比赛结果有关,因此,我们需要分不同的情况来讨论,然后把这几种情况所求的概率相加,才能得到莱斯特城队夺冠的概率,这就要用到概率论中的“全概率公式”。

设随机试验 E 共有 n 种可能的结果 $A_1, A_2, \cdots, A_n$，这些结果两两不可能同时出现，那么，任一随机事件 B 的概率满足：

$$P(B) = P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2) + \cdots + P(B \mid A_n) \cdot P(A_n)$$

这就是全概率公式。它隐含的思想正是我们在数学课上常用的“分情况讨论”，只不过，这里要求我们一定要把所有情况都列举全，而且不同的情况之间不能有交叉重叠。

在莱斯特城队登场前，我们先来热身一下。

请问，抛掷一枚硬币两次，出现至少一次正面的概率是多少？

有些读者会马上想到计算两次都是反面的概率，然后用 1 减去这个概率，这是个很聪明的想法，但在这里，我们要对全概率公式进行刻意练习。设

随机事件 A_1：第一次抛硬币出现正面；

随机事件 A_2：第一次抛硬币出现反面；

随机事件 B_1：第二次抛硬币出现正面；

随机事件 B_2：第二次抛硬币出现反面；

随机事件 C：两次至少出现一次正面。

根据全概率公式：

$$\begin{aligned} P(C) &= P(C \mid A_1) \cdot P(A_1) + P(C \mid A_2) \cdot P(A_2) \\ &= 1 \cdot P(A_1) + P(B_1) \cdot P(A_2) \\ &= 1 \times 1/2 + 1/2 \times 1/2 \\ &= 3/4 \end{aligned}$$

至少出现一次正面的概率是 3/4。

接下来，我们就用全概率公式来算一算莱斯特城队夺冠的概率。为了简化计算过程，我们仅用积分来度量莱斯特城队夺冠的可能性。过去 17 个赛季，英超冠军的最低积分为 79 分，2000 年之后，英超冠军的平均积分更是高达 87.5 分，就本赛季目前的积分情况，“低分冠军”似乎已成定局。虽然莱斯特城队现在领先优势不小，但是，“永远不要低估一颗冠军的心”，那些苦苦追赶的队伍有可能在最后 7 轮变身疯狂的抢分机器。因此，老帅拉涅利为球队定下了 79 分的目标，他认为，如果莱斯特城队在赛季结束时的积分能够达到甚至超过 79 分，便一定能夺冠。我们也以 79 分为标准，来计算莱斯特城队夺冠的

概率。

莱斯特城队夺冠的概率等价于莱斯特城队获得不低于 79 分的概率。31 轮过后，莱斯特城队积 66 分，距离 79 分还有 13 分。设

随机事件 A：莱斯特城队获得至少 13 个积分。

根据全概率公式：

$$
\begin{aligned}
P(A) = & P(A \mid \text{莱斯特城第 32 轮取胜}) \cdot P(\text{莱斯特城第 32 轮取胜}) + \\
& P(A \mid \text{莱斯特城第 32 轮打平}) \cdot P(\text{莱斯特城第 32 轮打平}) + \\
& P(A \mid \text{莱斯特城第 32 轮告负}) \cdot P(\text{莱斯特城第 32 轮告负}) \\
= & P(\text{莱斯特城后 6 轮取得至少 10 分}) \cdot P(\text{莱斯特城第 32 轮取胜}) + \\
& P(\text{莱斯特城后 6 轮取得至少 12 分}) \cdot P(\text{莱斯特城第 32 轮打平}) + \\
& P(\text{莱斯特城后 6 轮取得至少 13 分}) \cdot P(\text{莱斯特城第 32 轮告负})
\end{aligned}
$$

然后，我们还可以用全概率公式来计算 P(莱斯特城后 6 轮取得至少 10 分)、P(莱斯特城后 6 轮取得至少 12 分)和 P(莱斯特城后 6 轮取得至少 13 分)，按照同样的思路继续分解下去，直到最后一轮比赛。对于每一场比赛，我们要估计出莱斯特城队获胜的概率，然后将这些概率代入全概率公式中，便可以求得 $P(A)$。

我知道，我食言了，我没有算出莱斯特城队的夺冠概率，其实，我本就没打算真正去计算这个概率，毕竟，我们已经学习到了全概率公式的用法，这就足够了，至于莱斯特城队能否夺冠，我们只需要重温老帅拉涅利的那句话就可以了——“莱斯特能夺冠吗？我不知道，但能被问这个问题就足够美妙了。在这个金钱衡量一切的时代，我们给了每个人希望。”

第 2 章

随 机 变 量

导语：骰子是世人皆知的赌博道具。这个小小的赌博道具，对概率思想的启蒙做出了不可磨灭的贡献，伽利略、帕斯卡、费马等数学家从骰子的研究中发现了随机事件的数学本质，它就是随机变量。

2.1 随机变量：骰子游戏

骰子，俗称色子，是全世界都熟知的赌博道具。骰子的历史可以追溯到古巴比伦、古埃及时期，在中国古代的赌场里，也是赌博道具的不二之选。你可不能小看这小小的骰子，它对概率思想的启蒙做出了不可磨灭的贡献。

文艺复兴时期，意大利学者吉罗拉莫·卡尔达诺曾撰文研究骰子原理："在下注之前，你需要知道所有可能的结果，然后对比一下输赢的结果各有多少种，再按照这个比例去设置奖金，这样才能确保赌局的公平。"这大概是"概率思想"最早的启蒙，在当时是相当有革命性的思想。其后，著名的物理学家伽利略也对赌博中的数学原理产生了兴趣，并撰写了《骰子的研究》一书，在书中，他开创性的研究了掷多个骰子时可能出现的点数，以及这些点数会在怎样

的情况下出现。在那之后，赌博中的数学问题引起了很多学者的思考和讨论，其中包括著名数学家帕斯卡和费马。

我们回到过去，一起来看一看在概率论尚未建立时，聪明人是怎么利用骰子赚钱的。

掷骰子游戏

据资料记载，一个化名莫雷的赌徒曾经靠一个骰子游戏赚了很多钱，游戏的玩法是：连续掷骰子四次，如果出现至少一个六点，则莫雷赢；反之，莫雷输。要弄清楚莫雷为什么总是赢，就要计算一下双方赢的概率。要计算掷四次至少出现一个六点的概率，可以用逆向思维，计算掷四次没有任何一次出现六点的概率，再用 1 减去算出的概率即可，由于每次掷骰子都是彼此独立的，因此：

P(莫雷赢) ＝1－P(掷四次没有任何一次出现六点)

＝1－P(第一次没出现六点)×P(第二次没出现六点)×P(第三次没出现六点)×P(第四次没出现六点)

＝1－(5/6)×(5/6)×(5/6)×(5/6)

＝0.518

相对的，P(莫雷输)＝(5/6)×(5/6)×(5/6)×(5/6)＝0.482

莫雷赢得赌局的概率总是大于对手，所以莫雷可以靠这个赌局赚到钱，对吗？

不对！因为赌徒赚的可不是概率，是真金白银，我们忘记了赌局上最重要的东西——筹码。在莫雷的赌局中，双方的筹码是对等的，假定为“一两黄金”，也就是说，莫雷和对手各自拿出一两黄金作为筹码，如果出现了六点，莫雷拿走对手的一两黄金，如果没出现六点，莫雷将一两黄金送给对手。如表 2-1 所示，我们设定了一个关联关系——赌局结果与莫雷赢得的筹码之间的关联，莫雷赢得一两黄金的概率是 0.518，莫雷输掉一两黄金的概率是 0.482，如果将筹码的单位去掉，便可以表示成“＋1”对应的概率是 0.518，“－1”对应的概率是 0.482。

在概率论中，莫雷赢得的筹码就是一个随机变量。

表 2-1　莫雷赌局的结果

赌局结果	概率	莫雷赢得的筹码
莫雷赢	0.518	+1(赢得一两黄金)
莫雷输	0.482	−1(输掉一两黄金)

随机变量

假设随机试验有若干个可能的结果 $A_1, A_2, \cdots, A_n$,如果变量 X 满足:$A_1, A_2, \cdots, A_n$ 每一个都对应 X 的一个数值,那么,X 就称为随机变量。

上面的例子中,赌局是随机试验,赌局有两种可能的结果 A_1:莫雷赢,A_2:莫雷输,莫雷赢得的筹码是变量 X,A_1 对应 $X=+1$,A_2 对应 $X=-1$,所以,X 是一个随机变量。也就是说,随机试验的每一个结果都对应 X 的一个值。

一个随机试验可以包含不止一个随机变量,我们仍以骰子游戏为例。

小红、小黄和小蓝三个小朋友玩骰子游戏,规则是:扔一次骰子,出现一点或二点,小红赢;出现三点或四点,小黄赢;出现五点或六点,小蓝赢。游戏开始时,三个小朋友各自有五块泡泡糖,每局的赌注是一人一块泡泡糖,赌局一直进行到有人输光为止。

骰子的每个点数出现的概率都是 1/6,游戏中有三位小朋友,可以设定三个随机变量,分别是:

随机变量 X:小红一局赢得的泡泡糖数量;

随机变量 Y:小黄一局赢得的泡泡糖数量;

随机变量 Z:小蓝一局赢得的泡泡糖数量。

我们把游戏结果和随机变量一一列出,如表 2-2 所示。

表 2-2　骰子游戏的结果与随机变量

游戏结果	概率	X(小红)	Y(小黄)	Z(小蓝)
一点	1/6	+2	−1	−1
二点	1/6	+2	−1	−1
三点	1/6	−1	+2	−1
四点	1/6	−1	+2	−1
五点	1/6	−1	−1	+2
六点	1/6	−1	−1	+2

离散与连续

有这样一串数字,“1 2 3 1,1 2 3 1,3 4 5,3 4 5,5 6 5 4,3 1,5 6 5 4,3 1,2 5 1 0,2 5 1 0”,你能发现这串数字的奥秘吗?

也许有些读者一眼就看穿了我的把戏,但我还是不想现在就公布答案,我们先来讨论随机变量的两个平行世界——离散和连续。

现在,环顾你的四周,你能看到什么?你的手、这本书、手机、绿植等,这是我们看到的世界——宏观世界,这个世界里的东西总是可以用计数的,比如,你有两只手,你的手上有一本书,你有一部手机,手机里有两张 SIM 卡,你的绿植又生出了一片新叶。可是,世界并非全部如你所见。你一定记得,多年前的生物课上,当你第一次从显微镜里看到一团蠕动的细胞时,是多么的惊讶和好奇,那仿佛是另一个世界!科学告诉我们,显微镜下的细胞与我们看到的绿叶身处同一个世界,只不过它们太微小了,我们看不到。我们常把肉眼看到的世界称为宏观世界,把那个看不见的世界称为微观世界。

在数学世界里,也有宏观与微观的划分。我们从小学习的四则运算、一元二次方程等都是“宏观世界”的数学语言,直到我们遇上那几个让人抓狂的符号——“$\int$,Δ,∂”。从此,我们进入了数学的“微观世界”,那些简单的四则运算在“微观世界”里统统变了模样,它们演化成全新的运算规划——微积分。微微分扩张了概率论的疆域,随机变量不再只是赌徒的筹码,它也可以是时间、温度,于是,随机变量便自然地划分为两类——离散与连续。

离散随机变量,指的是随机变量的取值是有限的或可列无限个。比如,小红一局赢得的泡泡糖数量只有两个可能的取值;又如,一个把所有正整数都刻在上面的骰子,这个骰子掷出的点数可能是任何一个正整数,这就是“可列无限个”的离散随机变量。

连续随机变量,指的是随机变量的取值有无限多个并且不可列出。当我们把时间、温度、空间等无法一一罗列出来的指标作为随机变量的时候,连续随机变量就出现了。

有关离散随机变量和连续随机变量的讨论才刚刚开始,在后续章节中,我们会认识很多常用的随机变量,它们有些是离散的,有些是连续的,但无一例

外地都是概率论的重要成员。有关离散和连续的关系，我想了很久，想到了一个比喻：音符与音乐。一首曲子，曲谱只是一个个“离散”的数字，没有规律，没有内涵，但当这曲谱被演奏出来时，“离散”的数字化为“连续”的乐音，这乐音弥散在空中，让你陶醉，而你早已忘却了那一个个分离的音符，这就是“离散”与“连续”的完美结合。

最后，我要告诉你，那一串貌似神秘的数字其实是一首歌的乐谱，歌名是——《两只老虎》。

2.2　期望与方差：百变骰子

在当下的信息时代，人人生产信息，人人分享信息，我们忽然意识到，最稀缺的资源早已不是信息，而是人们的注意力。无论是一篇网文、一幅漫画，还是一部电影，引起人们注意的不二法则就是——简洁明确的特征。网文要有充满悬疑的话题；漫画要有个性鲜明的画风；电影则最好有一两个大牌明星，一切都要有特征，没有特征，便会沦为平庸。

在概率论的世界里，随机变量也像网文、漫画和电影一样需要特征，这些特征应该能够反映一个随机变量的本质，这些特征主要有两个，一个叫期望；另一个叫方差。

假定有四个不同的骰子，如图 2-1 所示，这四个骰子会带领我们认识期望和方差。

期望

期望是随机变量的第一个特征，它类似于我们常说的平均值，但又不是简单的求和平均。我们沿用上一节的例子来说明什么是期望。

还记得莫雷的骰子赌局吗？表 2-3 列出了莫雷赌局所有可能出现的结果，随机变量 X 表示莫雷赢得的筹码。根据 X 的取值和对应的概率，可以计算出 X 的期望：

$$E(X) = 0.518 \times (+1) + 0.482 \times (-1) = 0.036$$

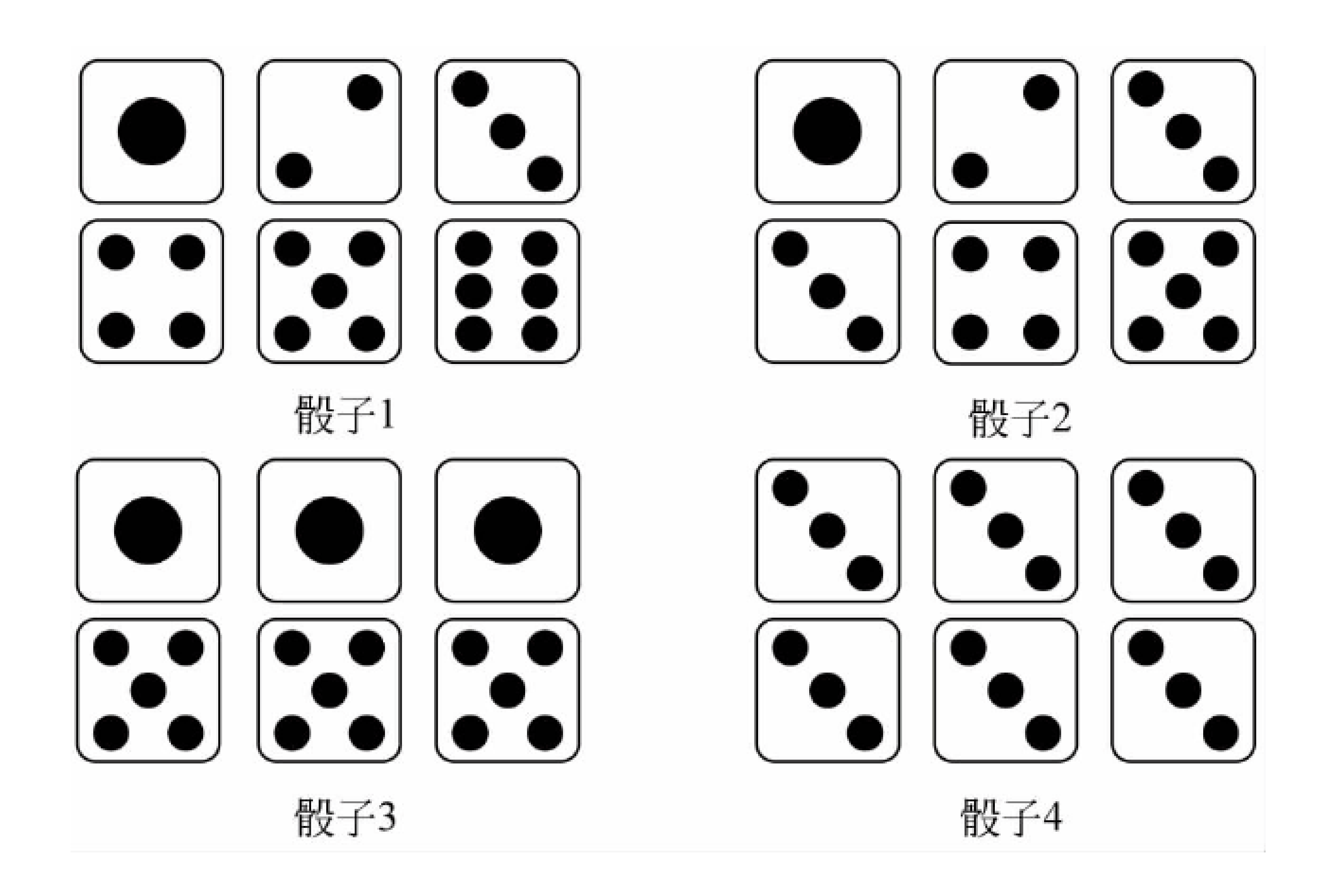


图 2-1 四个骰子游戏

由此，我们可以得到这样的结论：莫雷每一局所赢筹码的期望是 0.036 两黄金。

表 2-3 莫雷赌局的结果

赌局结果	概率	X（莫雷赢得的筹码）
莫雷赢	0.518	+1（赢得一两黄金）
莫雷输	0.482	−1（输掉一两黄金）

数学期望，简称期望，是随机变量的所有取值以对应概率为权重的加权求和。换言之，随机变量的每一个取值乘以它对应的概率，再相加求和，就得到了随机变量的期望。

设随机变量 X 有 n 个取值，分别是 $x_1, x_2, \cdots, x_n$，对应的概率分别是 $p_1, p_2, \cdots, p_n$，那么 X 的期望 $E(X)$ 是：

$$E(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n$$

这里需要说明，上一节我们提到过，随机变量分为离散和连续两种，由于连续性随机变量的计算涉及微积分，超出了本书的讨论范围，所以，本章只讨论离散随机变量。

下面，我们通过两个骰子游戏进一步理解期望。

- **骰子游戏 1**

掷骰子一次，随机变量 X 是掷出的点数，计算 X 的期望。

我们如法炮制，列出 X 的取值和对应的概率，如表 2-4 所示。由此可以求

得期望：

$$
\begin{aligned}
E(X) &= (1/6)\times 1+(1/6)\times 2+(1/6)\times 3+(1/6)\times 4+ \\
&\quad (1/6)\times 5+(1/6)\times 6 \\
&= 3.5
\end{aligned}
$$

表 2-4　骰子游戏 1 中随机变量取值和概率

游戏结果	概率	X(点数)
一点	1/6	1
二点	1/6	2
三点	1/6	3
四点	1/6	4
五点	1/6	5
六点	1/6	6

这个骰子的点数期望是 3.5，可是，骰子上可没有 3.5 这个点数，期望值是 3.5 代表了什么呢?

带着这个疑问，我们换一个骰子，把原来的六点改成三点，重新来过。

- **骰子游戏 2**

掷骰子一次，随机变量 X 是掷出的点数，计算 X 的期望。

根据表 2-5，可以求得期望：

$$
\begin{aligned}
E(X) &= (1/6)\times 1+(1/6)\times 2+(1/6)\times 3+(1/6)\times 3+ \\
&\quad (1/6)\times 4+(1/6)\times 5 \\
&= 3
\end{aligned}
$$

表 2-5　骰子游戏 2 中随机变量取值和概率

游戏结果	概率	X(点数)
一点	1/6	1
二点	1/6	2
三点	1/6	3
三点	1/6	3
四点	1/6	4
五点	1/6	5

这一次，点数的期望值是三点，刚好是 X 可能出现的点数，似乎是一个有意义的结果。可是，意义在哪里？难道反复抛掷骰子 B，最终就会一直出现三点吗？显然不是。

读者可以自己设计几个骰子，算一算它们的点数期望，看看期望和点数之间是不是存在联系。最终我们会发现，期望并不一定是随机变量的某一个值，期望可以是任何数值，即使它刚好与随机变量的某个取值相同，也与这个取值没有任何关系。期望只是随机变量的一个特征值，它就像一个球体的“球心”，随机变量的取值好比球体内的点，这些点分布在球心周围，甚至就是球心本身。因此，用期望来描述随机变量，就好像用球心来描述一个球体。但是球心不足以描述球体的全部特征，球体还有另一个特征——“半径”，随机变量的另一个特征“方差”正是用来描述“半径”的。

方差

我们继续做骰子游戏。

- **骰子游戏 3**

如图 2-1 所示，骰子 3 有六个面，却只有两个点数——一点和五点，表 2-6 列出了随机变量 X 的取值和概率，由此可以求得期望：

$$
\begin{aligned}
E(X) &= (1/6)\times 1+(1/6)\times 1+(1/6)\times 1+(1/6)\times 5+\\
&\quad (1/6)\times 5+(1/6)\times 5\\
&=3
\end{aligned}
$$

骰子 3 的点数期望与骰子 2 一样，可是，这两个骰子明显是不同的，这时我们需要用方差来区分这两个骰子。

表 2-6　骰子游戏 2 中随机变量取值和概率

游戏结果	概率	X（点数）
一点	1/6	1
一点	1/6	1
一点	1/6	1
五点	1/6	5
五点	1/6	5
五点	1/6	5

方差是随机变量取值与期望之差的平方，以对应概率为权重的加权求和。换言之，随机变量的每一个取值减去期望，求平方，再乘以它对应的概率，最后

求和,就得到了随机变量的期望。

标准差是方差的平方根,是与期望具有可比性的一个特征值。

设随机变量 X 有 n 个取值,分别是 $x_1,x_2,\cdots,x_n$,对应的概率分别是 p_1,$p_2,\cdots,p_n$,那么随机变量 X 的方差 $Var(X)$和标准差 $\sigma(X)$分别是

$$Var(X) = p_1 \cdot [x_1 - E(X)]^2 + p_2 \cdot [x_2 - E(X)]^2 + \cdots + p_n \cdot [x_n - E(X)]^2$$

$$\sigma(X) = \sqrt{Var(X)}$$

方差和标准差总是在一起使用,用来表示随机变量偏离期望的程度,偏离的程度越大,方差和标准差也越大,反之则越小。

以骰子 2 和骰子 3 为例,前面已经计算过,它们的点数期望都是 3,我们来计算方差和标准差。

骰子 2 的点数的方差是:

$$Var(X) = (1/6)\times(1-3)^2 + (1/6)\times(2-3)^2 + (1/6)\times(3-3)^2 + (1/6)\times(3-3)^2 + (1/6)\times(4-3)^2 + (1/6)\times(5-3)^2 = 1.67$$

骰子 2 的点数的标准差是:

$$\sigma(X) = \sqrt{1.67} \approx 1.29$$

骰子 3 的点数的方差是:

$$Var(X) = (1/6)\times(1-3)^2 + (1/6)\times(1-3)^2 + (1/6)\times(1-3)^2 + (1/6)\times(5-3)^2 + (1/6)\times(5-3)^2 + (1/6)\times(5-3)^2 = 4$$

骰子 3 的点数的标准差是:

$$\sigma(X) = \sqrt{4} = 2$$

很明显,骰子 3 的点数方差大于骰子 2 的点数方差,这说明骰子 3 的点数距离期望值更“远”一些,或者说,骰子 3 的点数更加分散,这一点从表 2-5 和表 2-6 中也可以看出。如果点数距离期望值非常近会怎样呢?

- **骰子游戏 4**

如图 2-1 所示,骰子 4 有六个面,每个面都是三点,表 2-7 列出了随机变量 X 的取值和概率,由此可以求得期望:

$$E(X)=(1/6)\times 3+(1/6)\times 3+(1/6)\times 3+(1/6)\times 3+(1/6)\times 3+(1/6)\times 3=3$$

方差：

$$Var(X)=(1/6)\times(3-3)^2+(1/6)\times(3-3)^2+(1/6)\times(3-3)^2+(1/6)\times(3-3)^2+(1/6)\times(3-3)^2+(1/6)\times(3-3)^2=0$$

标准差自然也是0。

表2-7　骰子游戏4中随机变量取值和概率

游戏结果	概率	X(点数)
三点	1/6	3
三点	1/6	3
三点	1/6	3
三点	1/6	3
三点	1/6	3
三点	1/6	3

骰子游戏4是一个极限情况，即随机变量的每一个值都一样，这时，期望一定就是这个值，方差也一定是0——方差和标准差的最小值。事实上，这样的极端情况仅存在理论可能性，并无实际意义，骰子的所有点数都相同，又何谈随机变量和概率呢？

协方差与相关系数

两个随机变量 X 和 Y 组合起来构成的随机变量 (X,Y) 称为二维随机变量，二维随机变量的方差称为协方差。

以骰子1和骰子2为例，设随机变量 X 为骰子1的点数，随机变量 Y 为骰子2的点数，X 和 Y 组成一个二维随机变量 (X,Y)，(X,Y) 的概率分布如表2-8所示。X 和 Y 的协方差用 $Cov(X,Y)$ 表示，计算公式为

$$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$$

由此前的计算结果可知：

$$E(X) = 3.5$$

$$E(Y) = 3$$

由表 2-8 中的数据,可以计算得到 X 和 Y 的协方差为:

$$Cov(X,Y) = 0$$

计算出协方差,便可以进而计算出随机变量 X 和 Y 的相关系数ρ_{XY},相关系数的计算公式为

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

表 2-8　二维随机变量(X,Y)的概率分布(1)

Y \ X	1	2	3	4	5	6	P(Y=i)
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/18	1/18	1/18	1/18	1/18	1/18	1/3
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
P(X=i)	1/6	1/6	1/6	1/6	1/6	1/6	1

相关系数ρ_{XY}可以用来判断随机变量 X 和 Y 的线性相关关系,$\rho_{XY}=0$ 说明 X 和 Y 不存在线性相关关系,$\rho_{XY}\neq 0$ 说明 X 和 Y 存在线性相关关系。上述例子中,由于 $Cov(X,Y)$为 0,所以ρ_{XY}也为 0,这说明骰子 1 的点数和骰子 2 的点数没有线性相关关系。

表 2-9 是另一组二维随机变量的概率分布,这是由两个标准骰子的点数组合而成的二维随机变量,根据协方差和相关系数的定义,可以计算得到:

$$Cov(X,Y) = -2.92$$

$$\rho_{XY} = -1$$

这说明 X 和 Y 存在线性相关关系,观察表中数据可以看出,X 和 Y 的关系是 $Y=7-X$,这也验证了我们的结论的是正确的。

表 2-9　二维随机变量(X,Y)的概率分布(2)

Y \ X	1	2	3	4	5	6	P(Y=i)
6	1/6	0	0	0	0	0	1/6
5	0	1/6	0	0	0	0	1/6

续表

Y \ X	1	2	3	4	5	6	$P(Y=i)$
4	0	0	1/6	0	0	0	1/6
3	0	0	0	1/6	0	0	1/6
2	0	0	0	0	1/6	0	1/6
1	0	0	0	0	0	1/6	1/6
$P(X=i)$	1/6	1/6	1/6	1/6	1/6	1/6	1

表 2-10 是第三组二维随机变量的概率分布，根据协方差和相关系数的定义，可以计算得到：

$$Cov(X,Y) = 0$$

$$\rho_{XY} = 0$$

这说明 X 和 Y 不存在线性相关关系。观察表中数据可以看出，X 和 Y 的关系是 $Y=X^2$，也就是说，$\rho_{XY}=0$ 只能用于说明两个随机变量不存在线性相关关系，无法判断二者是否存在非线性相关关系，这一点读者一定要谨记。

表 2-10　二维随机变量（X,Y）的概率分布（3）

Y \ X	1	2	3	4	5	6	$P(Y=i)$
1	1/6	0	0	0	0	0	1/6
4	0	1/6	0	0	0	0	1/6
9	0	0	1/6	0	0	0	1/6
16	0	0	0	1/6	0	0	1/6
25	0	0	0	0	1/6	0	1/6
36	0	0	0	0	0	1/6	1/6
$P(X=i)$	1/6	1/6	1/6	1/6	1/6	1/6	1

2.3　大数定理：庄家的信条

全世界有这样四个地方，不宜久留，因为你一旦到了那里，就会急不可待地把自己手中的钱拱手送人，它们就是世界四大赌城——亚洲澳门、欧洲摩纳哥以及美国大西洋城和拉斯维加斯。

提起赌场，我们自然会想到很多荧幕上的经典桥段，“赌神”总是能够在最

危急的时刻祭出唯一一张制胜牌，不仅让恶人们输得体无完肤，还会抱得美人归。可是，现实中的赌场里，根本不存在什么“赌神”，每个人都只是一个玩家。如果你是一个赌场新手，你的运气总是会出奇的棒，你下注，赢钱，再下注，又赢了钱，你扫视周围的玩家，他们摇头、瘪嘴、抱怨，只有你在暗自叫好：哈哈，我赢了他们的钱！你开始产生“赌神”附体的幻觉，你继续下注，一盘又一盘，最后，所有人的钱都输光了——自然也包括你。

你问：钱都去哪儿了？

我答：钱被“庄家”赢走了。

你问：谁是庄家？怎么赢的？

我答：庄家就是赌场，是那个为你准备扑克牌和香槟的人，他虽然没出现在赌桌前，却悄无声息的赚到了钱，他的信条总是会护佑他，让他赚到钱。

你问：他的信条是什么？

我答：全世界庄家的共同信条正是概率论中最经典的理论——大数定理。

大数定理

在抛硬币的例子里，有一个重要的前提条件——硬币的正面与反面出现的概率各为 50%。你觉得这看起来一定是对的吗？科学不相信感觉，科学相信实验。

下面，请准备好一枚一角的硬币（因为一角的更轻），咱们一起来做抛硬币的实验。实验过程是：高高抛起硬币并接住，每抛一次，都把结果记录下来，正面的次数 X 和反面的次数 Y 分别记录。

抛到 10 次，结果是，正面 3 次，反面 7 次。

抛到 100 次时，结果是，正面 43 次，反面 57 次。

抛到 200 次时，结果是，正面 97 次，反面 103 次。

抛到 1 000 次时，结果是，正面 513 次，反面 487 次。

这个实验可以永远进行下去，实验的目的不是找到某一次抛掷，使得 X 和 Y 刚好相等，实验的目的是观察 X 和 Y 的变化趋势。因此，实验暂时只进行到 1 000 次。图 2-2 是根据抛掷过程绘制出的曲线，曲线代表的是正面所占的比

例，即 $X/(X+Y)$ 随抛掷次数的变化。

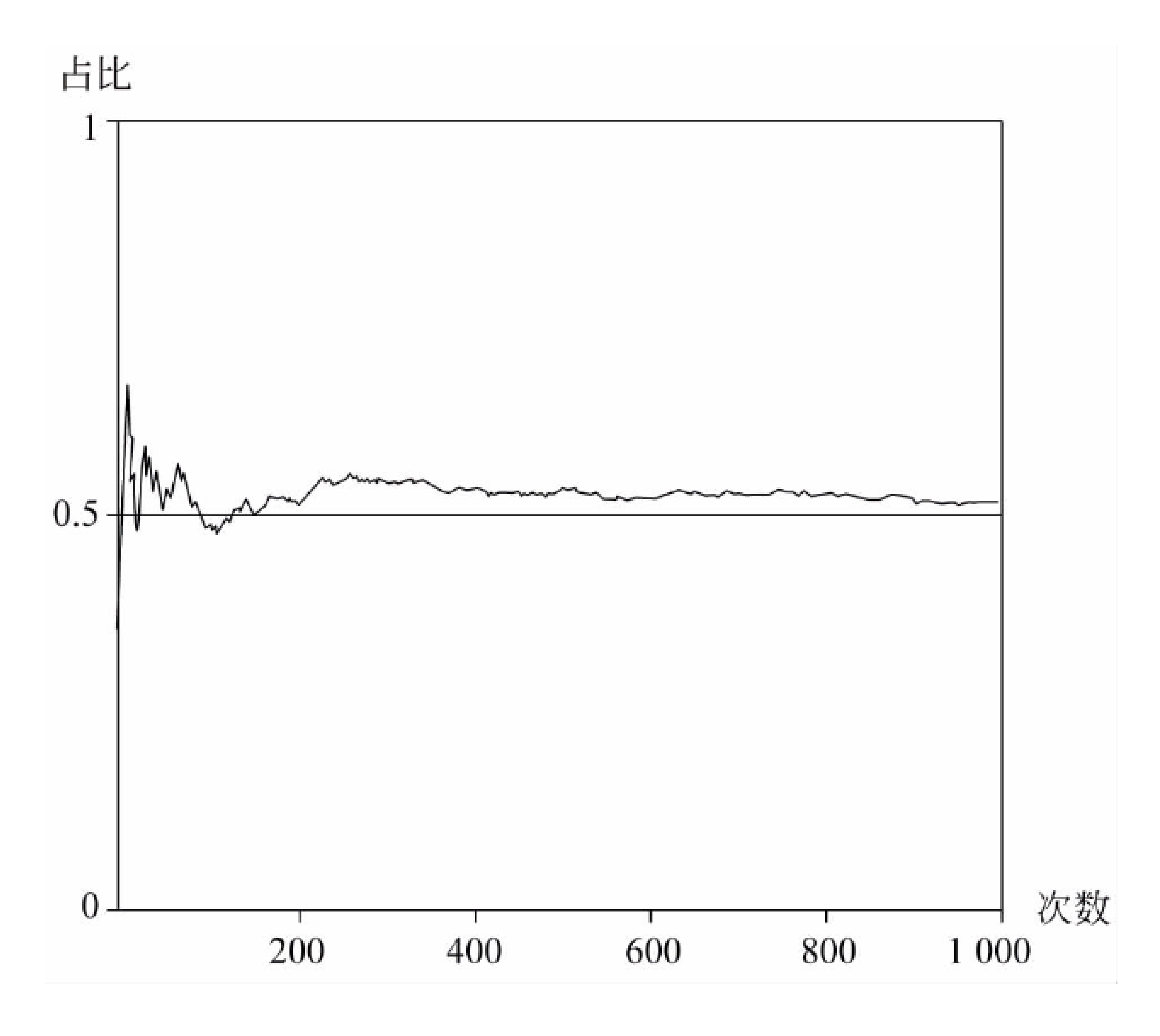


图 2-2　正面所占的比例随抛掷次数的变化

图中曲线呈现的特征是，当抛掷次数很少时，正面所占比例的变化幅度很大，并且与 0.5 的差值比较大，随着抛掷次数越来越多，正面所占的比例的变化幅度越来越小，而且一直围绕在 0.5 的周围。根据这条曲线，我们甚至可以预期，1 000 次之后的曲线还会在 0.5 周围徘徊，感兴趣的读者可以把实验继续做下去。

大数定理，指的是随机事件发生的频率会随着随机试验次数的不断增加趋向于它的概率，简单来说就是，试验次数越多，频率离概率越近，而且越稳定。在上面的实验中，随机事件是“抛硬币出现正面”，频率是“正面出现所占的比例 $X/(X+Y)$”，随着抛掷次数的增加，这个频率越发趋近概率值 0.5，大数定理像一只“看不见的手”，掌控着试验过程。

空手套利的庄家

我们回到赌场，坐回到赌桌前，看一看大数定理是怎么暗中帮助庄家赚到钱的。

我们要玩的是赌场里很流行的一个游戏——大转盘。游戏的道具是如

图 2-3 所示的大转盘,转盘上有 38 个格子,格子里填写了 1～36 的数字和两个特殊数字 0、00,玩家的下注方式有很多种,比如下注奇数,下注黑色格子的数字,或者下注某一个数字。这里需要特别说明的是,0 和 00 这两个数字不包含在任何赌注中,这两个数字是留给庄家的,也就是说,当转盘的指针最终指向 0 或 00 时,庄家赢得所有的筹码。

我们挑选赢的概率最大和最小的两种赌注。

赢的概率最小的赌注是下注某一个数字,当玩家下注某一个数字时,他赢的概率是 1⁄ 38,而此时庄家赢的概率是 2⁄ 38,很显然,玩家会输给庄家!

赢得概率最大的赌注是下注黑色(或红色)数字,当玩家下注黑色(或红色)数字时,他赢的概率是 18⁄ 38,这时,庄家赢的概率仍然是 2⁄ 38,很显然,玩家会战胜庄家!

很显然,上面的分析是错的!

因为玩家和庄家要赢的是筹码,可不是概率! 概率只是我们分析赌局的工具,玩家们真正关注的不是概率,而是所赢筹码的期望。为了计算所赢筹码的期望,我们首先要了解赌场里一个重要的常识——赔率。

赔率是赌场为每一个赌注设置的“赔钱比例”,比如,在 2015—2016 赛季英超联赛开始前,博彩公司为莱斯特城队开出的夺冠赔率是 1∶5 000,这个比例的含义是,玩家用 1 英镑下注莱斯特城队夺冠,如果莱斯特城队最终夺冠,博彩公司会付给玩家 5 000 英镑(含玩家下注的 1 英镑)。同时,阿森纳的夺冠赔率是 1∶3.5,即,下注阿森纳夺冠 1 英镑的玩家,即使赢了也只能得到 3.5 英镑。从这样的赔率可以看出,在英超联赛开始之前,博彩公司看好阿森纳夺冠,看衰莱斯特城队夺冠,这就是赔率的含义。

表 2-11 给出了大转盘中各类赌注的赔率,我们利用这些赔率来计算玩家和庄家所赢筹码的期望。

表 2-11　美式大转盘赔率

下 注 类 型	庄家开出的赔率
红色(或黑色)	1∶2
偶数(或奇数)	1∶2
1～18(或 19～36)	1∶2
任意 12 个数字	1∶3

续表

下 注 类 型	庄家开出的赔率
任意两行数字	1∶4
任意四个数字	1∶9
任意一行数字	1∶12
两个相邻数字	1∶18
一个数字	1∶36

假设玩家拿一个筹码下注某一个数字，他赢的概率是1/38，赢了可以得到35个筹码，输的概率是37/38，输了会输掉这一个筹码，所以玩家所赢筹码的期望是：

$$E(\text{玩家下注某个数字时，玩家所赢筹码}) = 1/38 \times 35 + 37/38 \times (-1) = -1/19 = -0.0526$$

与玩家相对的，庄家所赢筹码的期望是：

$$E(\text{玩家下注某个数字时，庄家所赢筹码}) = 1/38 \times (-35) + 37/38 \times (+1) = 1/19 = 0.0526$$

用同样的方法，可以计算出玩家下注黑色数字时，玩家和庄家所赢筹码的期望：

$$E(\text{玩家下注黑色数字时，玩家所赢筹码}) = 18/38 \times (+1) + 20/38 \times (-1) = -1/19 = -0.0526$$

$$E(\text{玩家下注黑色数字时，庄家所赢筹码}) = 18/38 \times (-1) + 20/38 \times (+1) = 1/19 = 0.0526$$

事实上，不论何种赌注，玩家所赢筹码的期望都是−0.052 6，庄家所赢筹码的期望都是0.052 6，读者们可以选择其他类型的赌注自行验证。

至此，我们终于看清了大转盘的本来面目，它是一个典型的“零和博弈”，庄家赢的筹码等于玩家输掉的筹码，平均意义上看，玩家每下注1个筹码，就会输掉0.052 6个筹码，同时庄家会赢得0.052 6个筹码。0.052 6看起来很微小，这正是庄家想要的效果，玩家就像温水中的青蛙，沉浸在赌局中，却不知

自己的钱正在像沙漏中的细沙一样,缓缓地流进了庄家的钱袋。

在这个赌局中,庄家要做到稳赚不赔,就要满足大数定理实现的条件:实验次数足够多。因此,庄家会想方设法地吸引玩家不停地玩下去,玩家越是沉迷于其中,庄家赚到的筹码也越多,这就是庄家空手套利的秘密。

大转盘示意图如图 2-3 所示。

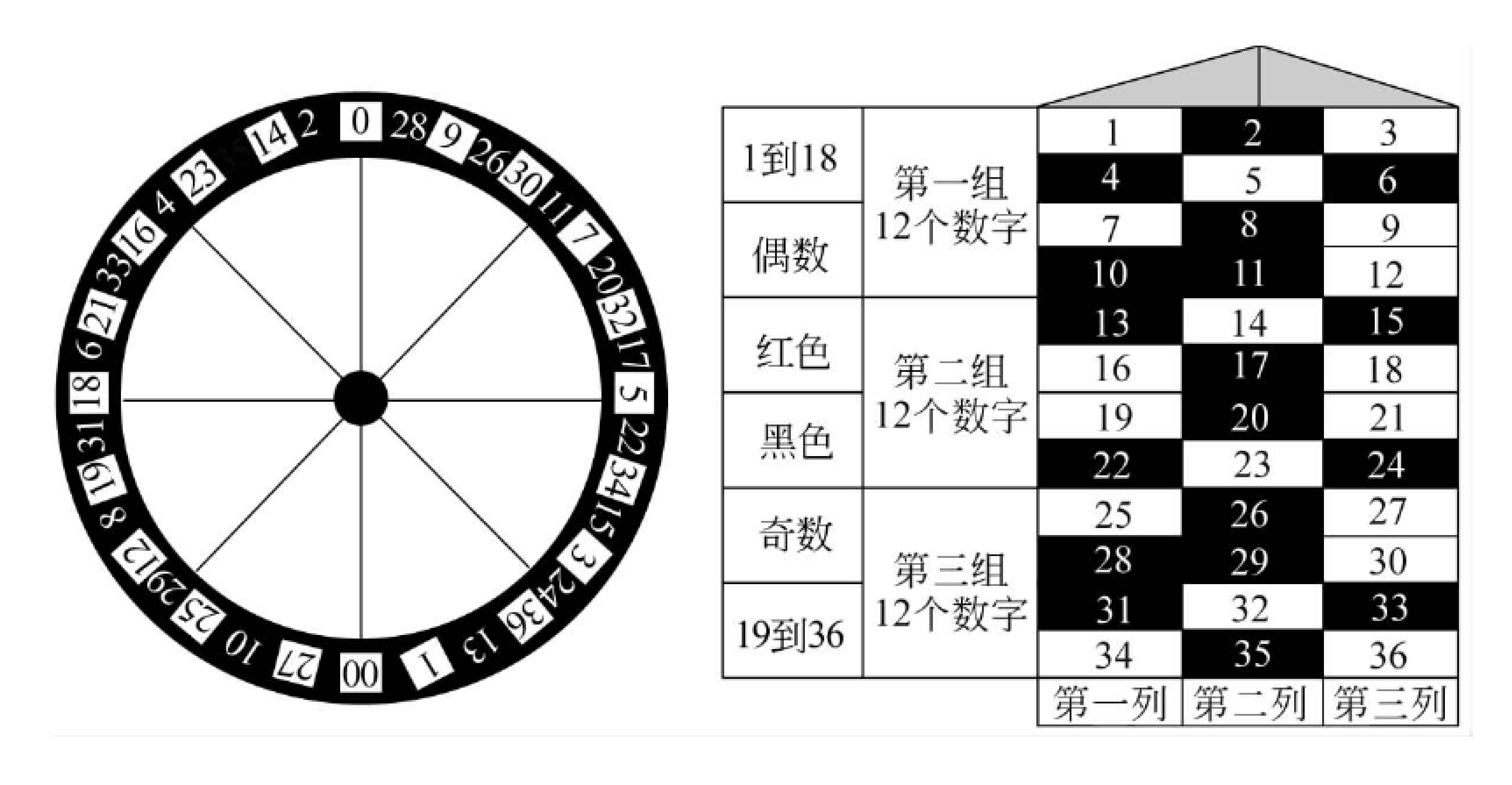


图 2-3　大转盘示意图

大数定理的误解

大数定理是概率论中最重要的定理,同时也是最容易被误解的定理。

在抛硬币试验中,我们发现,正面出现的频率随着抛掷次数的增加越来越接近 0.5 并且越来越稳定,这是大数定理作用于其中的结果,那么,这是否也说明,随着抛掷次数的增加,正面出现的次数和反面出现的次数也越来越接近呢?

在回答之前,我们需要分辨两个数学参量——相对频率和绝对频率。我们用 X 表示正面出现的次数,Y 表示反面出现的次数,N 表示抛掷次数。正面出现的相对频率是指 $X/(X+Y)$,正面出现的绝对频率是 X 本身,正面与反面出现次数的绝对频数差是 $X-Y$。我们已知,当 N 越来越大时,$X/(X+Y)$会趋近于 0.5 时,此时 $X-Y$ 是否也趋于 0 呢?我们通过实验来验证。

图 2-4 是抛掷硬币 1 000 次得到的两条曲线图,左图为相对频率 $X/(X+Y)$与抛掷次数 N 的关系曲线,右图为绝对频数差 $X-Y$ 与抛掷次数 N 的关系

曲线。右图中，随着 N 的增大，$X-Y$ 并没有越来越趋近于 0，仍然变化不定。通过这个反例，我们可以否定“正面出现次数与反面出现次数越来越接近”的说法。更加反直觉的结论是，X 与 Y 相等的概率会随着 N 的增加越来越小！这个结论会在“二项分布”一节中做出解释。

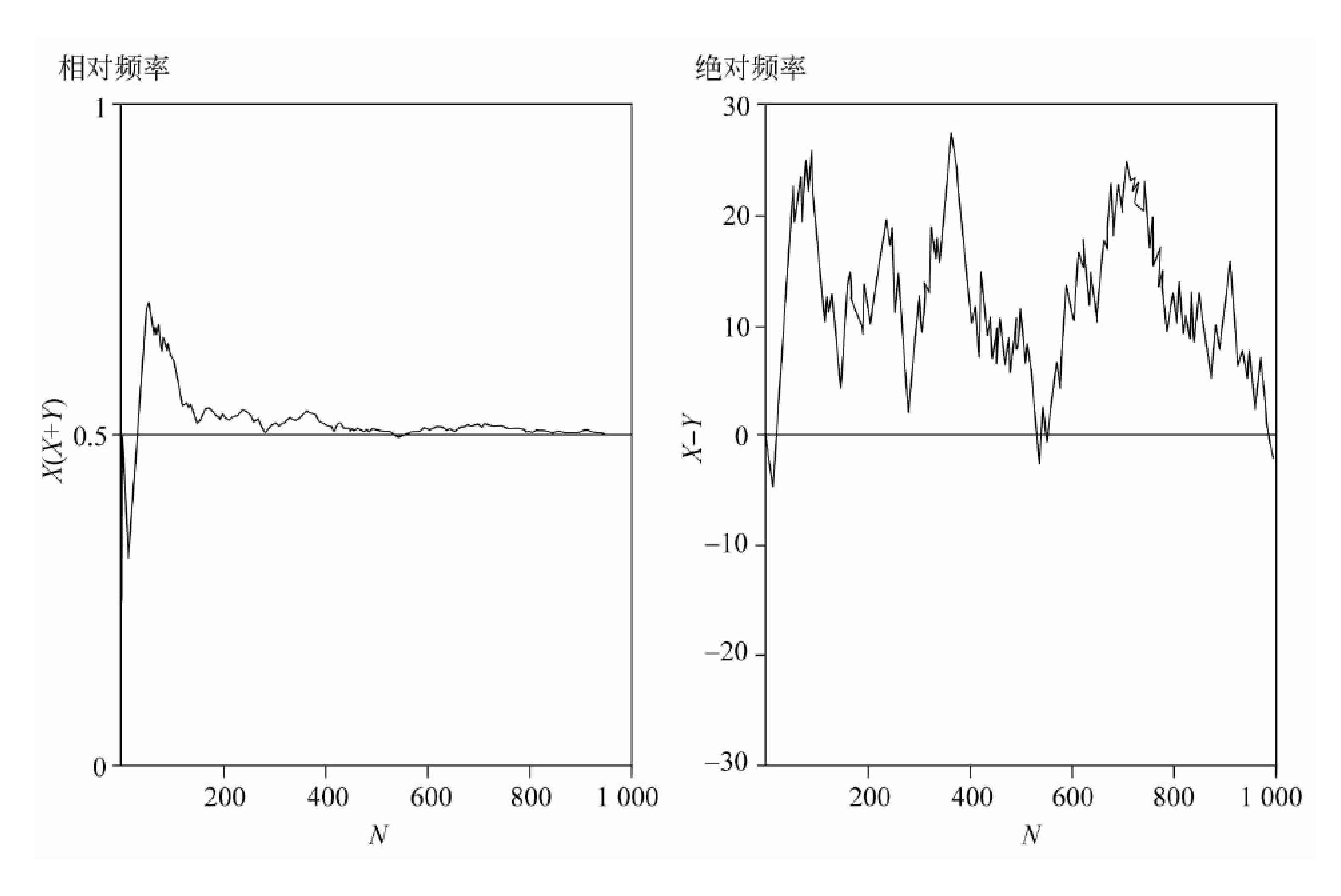


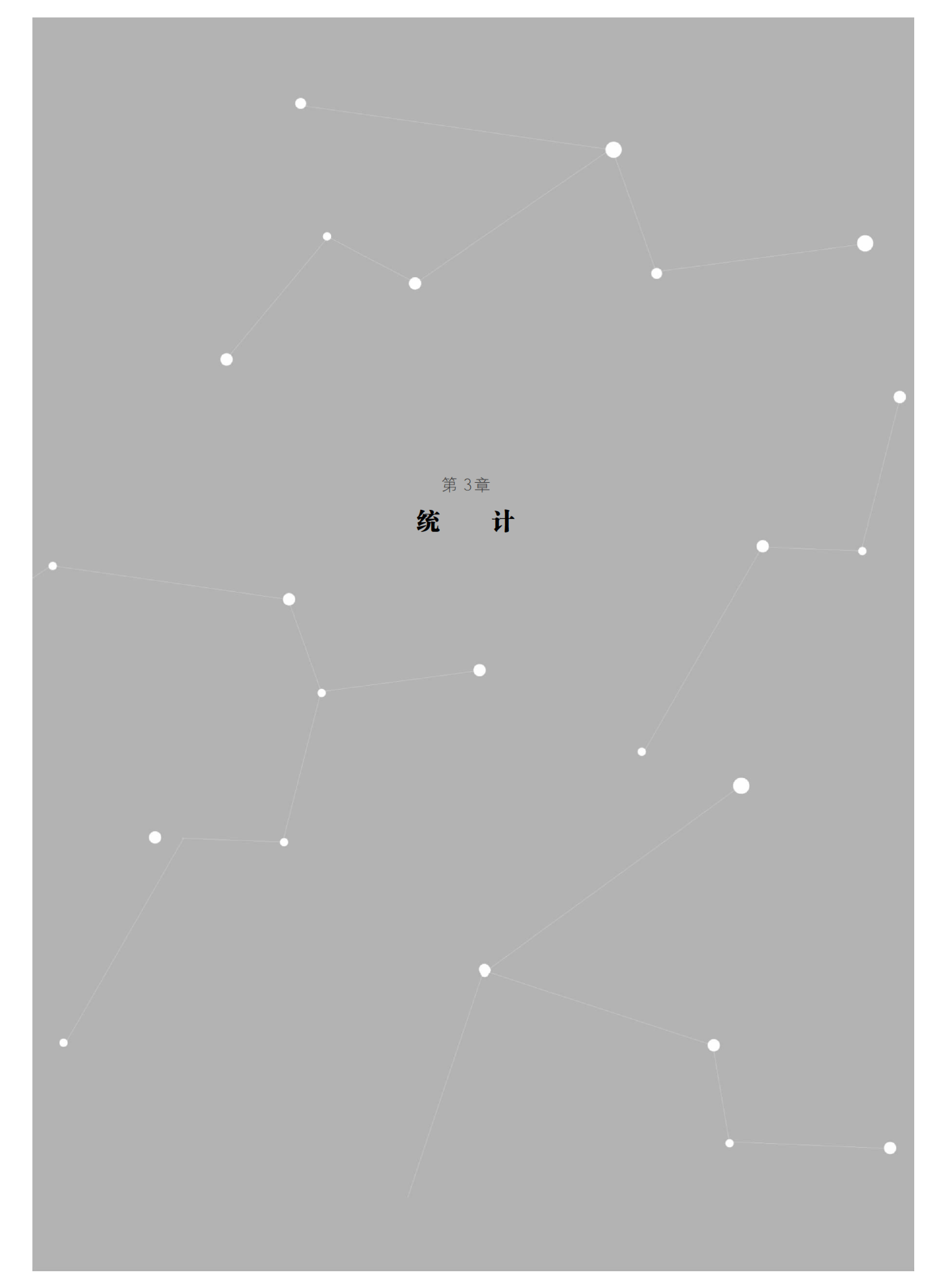
图 2-4　抛掷硬币 1 000 次的相对频数和绝对频数差

在很多赌博游戏中，玩家会对大数定理保有另一个误解：如果反复进行的试验偏向某些结果，那么后边的试验结果很可能会偏向其他结果。举个例子，如果抛硬币 10 次，正面出现了 7 次，反面出现了 3 次，下一次抛掷出现反面的概率会更大吗？我们已经学过独立事件，所以我们要相信，概率依然是 50%。可是，这似乎和大数定理矛盾，我们要弥补正面与反面的差值才能让正面出现的次数趋于 0.5，难道不是吗？

还真不是！事实上，要让概率趋近于 0.5，我们根本不需要弥补此前的不均衡。举一个极端的例子，假如接下来，每抛 10 次，都会出现 5 次正面、5 次反面，那么，抛掷 20 次时，正面出现的相对频率会从 0.7 下降到 0.6，再抛 10 次会下降到 0.57，再抛 10 次会下降到 0.55，以此类推，越来越趋近于 0.5。也就是说，只要硬币一直随机出现正反两面，大数定理依然成立，根本不需要刻意

弥补此前的空缺！从另一个角度来看，随着抛掷次数的逐渐增加，前 10 次的抛掷结果对相对频数的贡献越来越小。因此，我们并不需要弥补这个小小的缺口。

总之，大数定理只是在描述随机现象的规律，它只会告诉你长期的、平均的情况，它无法预测未来。



第 3 章

统　　计

导语：概率和统计像一对性格迥异的兄弟，概率是理想主义的“文艺青年”；统计是务实精干的“普通青年”。概率喜欢提出很多“假设”和“近似”；统计则只顾着搜集数据，分析数据，寻找数据中隐藏的秘密。

3.1 从样本到总体：管中窥豹

前面两章，我们学习了概率的基础知识，本章我们一起来认识概率的亲兄弟——统计。

如果说概率论像一个理想主义的“文艺青年”，统计学则是一个务实精干的“普通青年”，在统计学中没有那么多“假设”和“近似”，统计学研究实实在在的数据，从数据中发现规律，再利用规律指导我们的行动。因此，数据是统计学的基础。

在统计学中，数据被自然的分为两类：样本与总体。举个例子，假设味多美公司刚刚出品了一款巧克力慕斯蛋糕，为了检验这款蛋糕的受欢迎程度，味

多美在很多超市里举办免费试吃，并让试吃者填写一份简单的调查问卷。试吃活动进行了两周，收到了一万多份问卷。味多美整理分析了这些调查问卷的内容，针对不同年龄、不同性别的消费者各自进行了分析，发现年轻男性十分喜欢这款蛋糕，于是味多美决定，到中关村和理工科大学去推广这款蛋糕。在这个虚构的例子中，味多美公司想要测试新款蛋糕的受欢迎程度，如果它可以让所有消费者都试吃一次，那么它就可以从试吃结果中精确地找到喜欢这款蛋糕的人群，这么做成本高的离谱，显然无法实现。于是它退而求其次，挑选几个超市开展试吃活动，吸引一部分消费者来品尝，获得他们的反馈。从统计学的角度来看，“所有消费者的反馈”是总体，“部分消费者的反馈”是样本。

总体，是指一个试验中所有可能的观察值。这些观察值有时是有限多个，比如全校学生的身高；有时是无限多个，比如宇宙中的所有行星，统计学的目标是研究总体中包含的统计学规律。然而，总体往往难以全部获得，因此，我们从总体中抽取一部分观察值，通过研究它们的规律推理出总体的规律，这部分被抽取出来的观察值就是样本。从样本推测总体，正如管中窥豹，虽然只“可见一斑”，却依然要从这“一斑”推想出“全豹”。

数据会说谎

前面我们提到，数据是统计学的基础，要学习统计学，首先要学会正确地看待数据，有时数据是会说谎的。

有这样一个思想实验。很久很久以前，有一个原始人，住在现在的北京所在的地方。他每天早晨从山洞里跑出来，迎接日出，然后出去捕猎，直到太阳落山后，才跑回山洞里睡觉。一天又一天，太阳升起又落下，每天晚上入睡时，他都十分确信，明天早晨，太阳会照常升起。在另一个地方，一个特别寒冷的地方，也有一个原始人。他的头上一直悬着一个太阳，于是他以为，太阳会永远发光。忽然有一天，太阳消失了，消失得无影无踪，刺骨的寒冷夺去了他的生命。直到死去，他也不明白，太阳究竟去哪儿了。

两个原始人看到了同一个太阳，却对太阳的认识相去甚远。这个简单的思想实验告诉我们，样本的规律未必能代表总体的规律，你以为太阳升起落下是必然规律，是因为你没去过北极。

在统计学中，由片面的样本推理总体的规律往往会以偏概全，这种现象被称为“幸存者偏差”，更通俗的说法是——“死人不会说话”，第二次世界大战时期美国战斗机的故事正说明了这一点。

第二次世界大战时期，美英联军出动大量战斗机，对德国展开大规模空袭，但是德军强大的防空火力让美英联军遭受重创。为了对抗德军的防空火力，美英联军找来了飞机领域的多位专家，要求他们研究战斗机的受损情况，对飞机的设计制造提出改进意见。飞机专家们对执行任务归来的飞机进行了仔细地检查，发现几乎所有的飞机的机腹都伤痕累累，于是专家们建议，加固机腹。可是，美英联军最终没有采纳飞机专家的意见，反而加强了对机翼的防护。这是因为，国防部的一位统计学家认为，能够幸运返航的飞机，机翼大多完好无损，这说明，被击中机翼的飞机都坠落了，而仅被击中机腹的飞机却能够顺利返航，说明机腹不是要害部位，不需要进行加固。因此，他建议美英联军加强对机翼的防护。

在上面的事例中，飞机学家由于缺少统计学知识，错把顺利返航的飞机与被击落的飞机混为一谈。他们把“顺利返航的飞机”作为样本，来推测总体的规律，恰恰掉入了“幸存者偏差”的陷阱中。反观统计学家，从总体出发来寻找规律，虽然他无法观察到被击落的飞机，但他观察顺利返航的飞机之后，推测出了被击落的飞机可能的受损情况，进而提出加固建议，是更合理的解题思路。这个例子除了提醒我们提防“幸存者偏差”之外，还告诉我们，弄清研究对象十分重要，被击落的飞机才是正确的研究对象。

另有一类数据也容易混淆视听，那就是“小概率事件”相关的数据。小概率事件是一些生活中非常稀有但切实发生的事件，最常听到的就是彩票中大奖和被雷劈。小概率事件的发生概率也是通过数据计算出来的，比如，要计算被雷劈中的概率，只需要用被雷劈中的人数除以总人口便可以得到，大约接近百万分之一。然而，小概率事件由于样本十分稀少，往往容易出现大幅波动，引起人们的误解。

马航 370 事故让空难再次发酵成一个热点话题，在民航领域，衡量民航安全的重要指标是致死事故率，它是指每一百万次航班中的致死事故总数。在 20 世纪后半叶，由英国和法国联合研制的协和式超音速客机是全世界最安全的客机，在 2000 年 7 月的空难发生前，协和式飞机共飞行了约八万次，从未发

生过致死事故，因此致死事故率为 0，与之同期的波音 737 飞机，飞行了约一亿五百万次，致死事故率为 0.41。然而，2000 年 7 月，协和式飞机不慎发生空难，仅仅这一次空难，使协和式飞机的致死事故率瞬间升至 12，一跃成为全球最危险的飞机！

另一个例子是谋杀率。谋杀率是衡量一个国家是否安全的重要指标，在任何一个长期稳定的国家，一年里发生的谋杀案都很少，在 13 亿人口的中国如此，在不足千人的梵蒂冈也是如此。梵蒂冈是全世界人口最少的独立主权国家，只有不足千人，由瑞士卫队保卫国家安全。多年来，梵蒂冈从未发生过谋杀案件，直到1998 年 5 月 4 日晚，瑞士卫队队长阿洛伊斯·埃斯特曼和妻子被枪杀。这一晚之后，梵蒂冈的谋杀率瞬间达到五百分之一，领跑全球谋杀率排行榜，成为全世界最不安全的国家。后来，梵蒂冈回归了宁静，谋杀率也重新降回零。

小概率事件总是很少发生，由数据计算出的发生概率是否有意义，值得质疑。很多时候，小概率事件的概率只是新闻媒体的噱头。从概率统计的角度来看，它只能告诉我们，这件事很少发生。

抽样

前面我们提到，从总体中抽取一部分可以获得样本。在统计学中，这个抽取的过程叫作抽样。

抽样有自己的方法，最简单、最常用的抽样方法是简单随机抽样，比如味多美可以随机挑选几个地方举办蛋糕试吃活动，并在活动过程中随机招揽路人来试吃。在试吃活动中，味多美的服务人员可以给参加试吃的人免费发放购物袋，这样他们就可以辨认出哪些人已经参加过试吃活动，不再招揽他们参加试吃，这就是不重复随机抽样。如果味多美放任所有人试吃，不做任何筛选和限制，就是重复随机抽样。

在简单随机抽样中，重复抽样和不重复抽样都是常见的抽样方式。比如，同样是福利彩票，33 选 7 的双色球采用的是不重复抽样，排列 3、排列 5 采用的是重复抽样。在进行数据抽样时，我们根据事件的需要选择抽样方式。

除了简单随机抽样，还有其他几种抽样方法。一个是分层抽样，仍以味多

美为例，服务人员可以分别邀请年轻女性、年轻男性和儿童参加试吃活动，也就是按照年龄和性别对人群分组，再进行抽样，这就是分层抽样，也可以理解为先分组再抽样；另一个是整群抽样，假定新款的蛋糕有草莓、樱桃和杧果三种配搭的水果，服务人员可以将蛋糕分装到不同的盒子中，每个盒子里放置草莓、樱桃和杧果蛋糕各一块，让消费者们整盒的进行试吃，这种抽样方法便于对比，从对比结果可以看出哪种口味更受欢迎。还有一些抽样方法，本书不再一一介绍，无论采用什么方法，我们的终极目标都是采集到能够代表总体的样本。

读到这里，想必读者会有这样的疑问：现在都是大数据时代了，还需要抽样吗？诚然，在互联网行业里，抽样的概念的确过时了，正如《大数据时代》一书所说："在大数据时代进行抽样分析就像在汽车时代骑马一样。"在互联网行业，样本几乎就是总体，谷歌、苹果和淘宝这些公司甚至不需要刻意的搜集数据，只需要利用互联网软件记录下人们在手机和计算机上的每一次触碰和点击，便完成了数据采集。但是互联网不能代表一切，很多数据并不能从互联网上搜集，比如前面例子中提到的试吃体验数据。所以，在互联网力所不及的领域，采用抽样的方法搜集数据仍是必要的。

3.2　频数、均值与中位数：致敬"黑曼巴"

2011 年 2 月，耐克公司推出了一部广告电影《科比就是黑曼巴》，NBA 球星科比・布莱恩特从此得到了一个新绰号——黑曼巴。黑曼巴蛇属于眼镜蛇科，生长于非洲草原和林地，是全世界最致命的毒蛇。除了剧毒，黑曼巴还拥有闪电般的速度，其短距离移动时速可达 16～20 公里，能在几分钟内杀死 13 个围捕者；黑曼巴喜欢独居，仿佛孤独是它的天性；黑曼巴十分贪婪，它会一口把猎物吞下，即使是最难消化的食物也会在几小时内消失。正如电影片名所说，科比就是黑曼巴，自从 18 岁加入 NBA 联盟起，科比就开始展现自己"黑曼巴"的天性，他突破速度极快，能够单场独得 81 分，但是球风偏独，常常被人诟病。不论怎样，当令人窒息的读秒阶段到来时，科比永远是执行绝杀球的不二人选，这时的科比就像剧毒的黑曼巴，随时会在红灯亮起前给予对手致命

一击。

“最接近神的球员”是科比的另一个绰号,“神”指的自然是“篮球之神”迈克尔·乔丹。在科比职业生涯的巅峰期,媒体和球迷们常常拿科比和乔丹做对比,他们会列出两人的各项技术统计,逐一对比,然后写出一篇“科比与乔丹,到底谁更强?”的软文。今年,科比将正式退役,全世界的篮球迷们都必须对他二十年的职业生涯表达敬意。接下来,我们抽取科比的部分统计数据,一起来学习三个常用的统计量——频数、均值和中位数。

频数

表 3-1 是科比 2008—2009 赛季常规赛的每场得分数据,下面我们一起来分析这组数据。

表 3-1　科比 2008—2009 赛季常规赛每场得分数据

23	32	22	26
16	26	38	37
33	28	30	23
27	28	25	28
23	41	61	11
27	36	36	21
20	26	26	28
29	27	19	19
21	31	34	30
24	40	37	14
29	26	10	17
24	39	30	25
12	21	39	30
35	36	28	20
23	19	36	18
28	33	22	22
32	29	29	33
23	28	49	32
20	20	31	16
28	18	23	16
18	11		

通过观察，我们可以找出最大值为 61，最小值为 10。我们想知道，科比的得分在最大值和最小值之间是如何分布的，这时我们需要制作一个频数分布表，绘制一张直方图。

我们将最小值到最大值之间划分为 6 个小范围，也称为 6 个区间，分别是 10～20、21～30、31～40、41～50、51～60、61～70，统计有多少个数据落在这 6 个区间内，并记录下来，便得到了如表 3-2 所示的频数分布表。

表 3-2　科比得分的频数分布表

分组	频数	相对频数	累积频数
10～20	19	0.232	19
21～30	40	0.488	59
31～40	20	0.244	79
41～50	2	0.024	81
51～60	0	0.000	81
61～70	1	0.012	82

表中的第一列是分组方式；第二列是频数，即每个区间里有多少个数据；第三列是相对频数，即频数除以数据总量；第四列是累积频数，即对频数进行累积计数。这张表格包含了数据分析的三个重要的思路：一是分类统计，体现在频数中，即把数据按照某种属性进行分类计数；二是相对数量，体现在相对频数中，相对频数的本质是将频数进行“归一化”，这样便于与其他数据进行对比；三是累积数量统计，体现在累积频数中，对数量进行累积统计便于我们观察出数量的变化规律，也便于我们快速找出低于或高于某些临界值的数据有多少，比如，从累积频数一列中，我们可以知道，低于 30 分的有 59 场，低于 40 分的有 79 场。

图 3-1 是科比得分数据的直方图，直方图与频数分布表相对应，是通过绘图的方式更直观地展现频数分布情况，直方图中每一个条形都代表一个分组，条形的高度代表频数。频数分布表和直方图是统计学中的常用图表，也是数据分析的第一步。

均值

平均值，简称均值，是最常用的统计量，计算方法是用总量除以数量。例

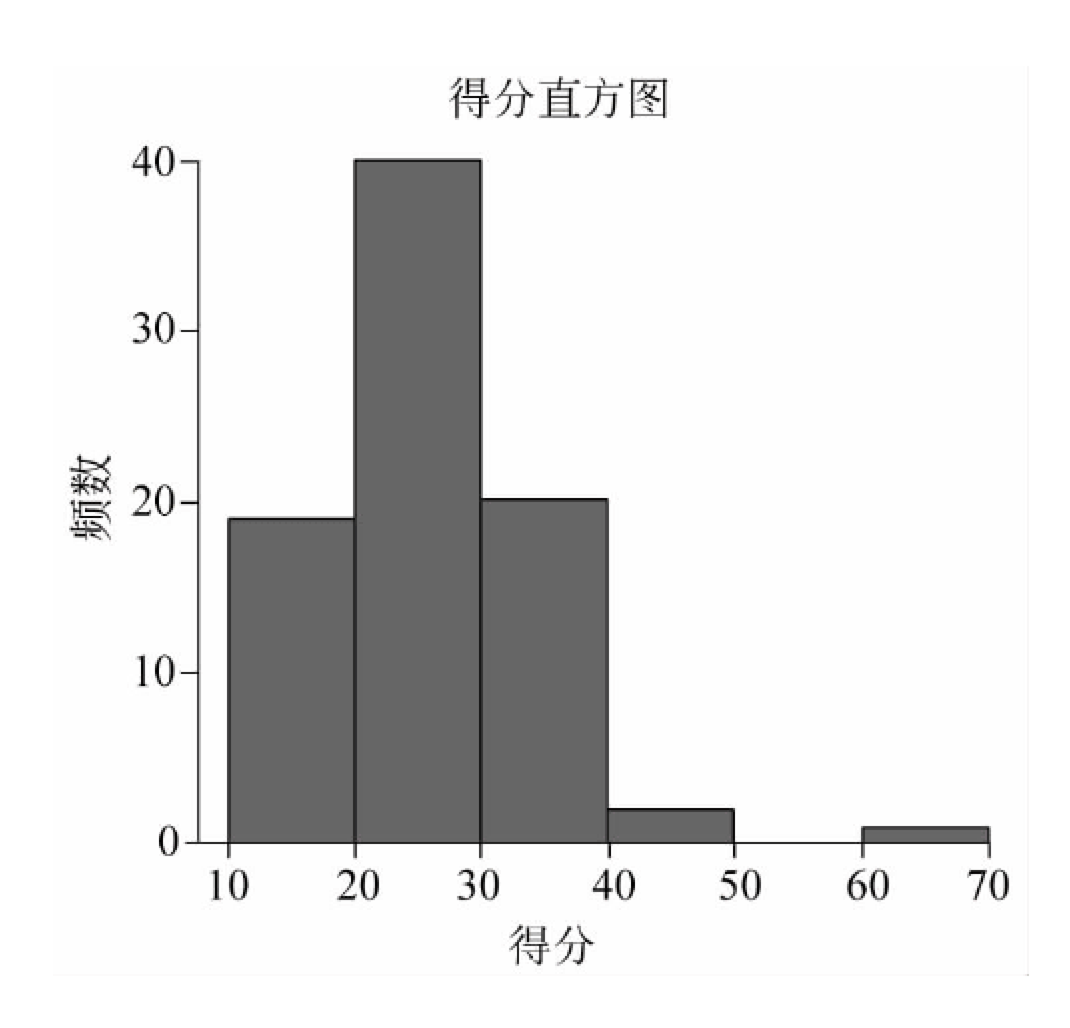


图 3-1　科比得分的直方图

如,2015 年我国的国内生产总值 GDP 为 67.67 万亿元,我国同期的人口总数约为13 亿,因此,人均 GDP 为 5.2 万元。表 3-3 是科比 2008—2009 赛季 82 场常规赛的各项技术统计,取出其中的一列数据,全部相加后除以 82,便可以计算出科比的场均技术统计。

表 3-3　科比 2008—2009 赛季常规赛技术统计

场次	得分	篮板	助攻	抢断	封盖	失误
1	23	11	5	1	0	5
2	16	8	3	2	0	5
3	33	4	3	2	0	1
4	27	2	3	2	2	2
5	23	3	3	2	2	4
6	27	4	1	1	0	1
7	20	4	6	0	0	2
8	29	5	6	4	1	3
9	21	5	6	3	2	1
10	24	5	3	0	1	2
11	29	4	2	2	0	1
12	24	4	6	3	1	3
13	12	6	4	1	0	0
14	35	6	5	1	1	3
15	23	5	7	1	0	3
16	28	7	2	1	0	1

续表

场次	得分	篮板	助攻	抢断	封盖	失误
17	32	6	4	2	0	4
18	23	7	7	0	0	2
19	20	5	8	2	0	4
20	28	4	3	0	0	2
21	18	7	3	1	0	5
22	32	7	3	0	1	4
23	26	3	5	2	0	2
24	28	7	6	4	0	1
25	28	3	3	0	0	5
26	41	8	3	0	0	1
27	36	4	3	2	0	5
28	26	6	4	0	0	0
29	27	9	5	1	0	4
30	31	3	4	4	0	2
31	40	7	4	2	1	2
32	26	2	3	1	0	4
33	39	4	7	0	1	4
34	21	5	5	2	0	3
35	36	7	13	1	0	5
36	19	2	7	0	0	4
37	33	7	4	0	0	2
38	29	7	10	0	1	4
39	28	13	11	2	0	6
40	20	6	12	1	1	5
41	18	10	12	1	0	4
42	11	4	5	0	0	0
43	22	4	3	1	1	1
44	38	8	5	1	0	2
45	30	8	5	0	0	2
46	25	1	7	3	0	1
47	61	0	3	0	1	2
48	36	9	5	2	0	0
49	26	10	5	1	2	3
50	19	3	2	1	0	2
51	34	7	1	0	0	3
52	37	4	4	4	0	4
53	10	4	2	0	2	4

续表

场次	得分	篮板	助攻	抢断	封盖	失误
54	30	3	9	3	2	2
55	39	5	5	1	1	2
56	28	6	7	1	0	6
57	36	4	5	1	0	4
58	22	4	8	3	0	1
59	29	8	2	0	1	2
60	49	11	2	1	1	1
61	31	2	2	1	0	0
62	23	2	4	1	0	1
63	26	3	3	1	1	1
64	37	5	6	4	2	4
65	23	4	6	0	1	2
66	28	8	5	1	1	1
67	11	5	5	2	0	5
68	21	6	2	2	0	5
69	28	3	7	5	1	3
70	19	3	5	4	2	0
71	30	8	7	2	0	1
72	14	1	9	3	1	3
73	17	8	4	1	0	4
74	25	2	2	3	0	1
75	30	8	4	3	0	4
76	20	1	7	2	0	3
77	18	4	5	1	1	2
78	22	5	4	2	0	0
79	33	3	2	1	0	2
80	32	5	2	0	0	3
81	16	7	4	2	1	1
82	16	1	5	2	0	1
场均	26.84	5.23	4.87	1.46	0.45	2.56

例如，我们用 X_1、X_2、…、X_{82} 分别表示科比 82 场比赛的篮板数，那么，场均篮板数 $\overline{X}$（读作 X 拔）为

$$\overline{X} = (X_1 + X_2 + \cdots + X_{82}) / 82 = (11 + 8 + \cdots + 1) / 82 = 5.23$$

在人均 *GDP* 和场均篮板数的例子中，我们计算的平均值都是“算术平均值”，以两个数 *A* 和 *B* 为例，算术平均值就是 $(A+B)/2$。统计学中还有其他

几种均值，分别是几何平均值、调和平均值和均方根值。

例如，股神巴菲特去年的资产增长了 50%，今年减少了 4%，那么，这两年的平均增长率就是 $\sqrt{1.5\times0.96}=1.2$，平均增长率是 20%，这就是几何平均值。又如，火车从北京到上海的平均时速是 200 公里/小时，从上海到北京的时速是300 公里/小时，那么，来回的平均时速是 2/(1/200+1/300)=240 公里/小时，这是调和平均值。均方根值的计算方法是 $\sqrt{(A^2+B^2)/2}$，在标准差的计算中会用到均方根值。在后文中，如无特殊说明，均值都是指算术平均值。

在统计学中，计算均值往往只是第一步，很多时候，我们会将不同的均值进行比较，这时，我们一定要小心“辛普森悖论”的陷阱。“辛普森悖论”是由英国统计学家辛普森发现的，这个悖论让我们更深刻的认识和修正了均值比较的方法。下面，我们以科比和乔丹的得分为例，来说明“辛普森悖论”。

表 3-4 是两组假想的得分数据，在 1996—1997 赛季中，科比由于肩伤只打了 17 场比赛，乔丹则打了 80 场，到了 1997—1998 赛季，两人都保持健康，科比更是 82 场常规赛保持全勤，观察两人的场均得分可以发现，这两个赛季乔丹的场均得分都高于科比，毕竟那时的科比还是个毛头小子，乔丹则处在职业生涯最后的辉煌时期。我们的问题是，两个赛季平均下来，谁的场均得分更多？

表 3-4　乔丹和科比两个赛季的得分假设值

	球员	总得分	场次	场均得分
1996—1997 赛季	乔丹	2 182	80	27.3
	科比	440	17	25.9
1997—1998 赛季	乔丹	2 832	80	35.4
	科比	2 870	82	35.0

每个赛季都是乔丹得分更高，难道两个赛季加在一起，乔丹还会比科比低吗？事实告诉我们，乔丹还真比科比低。如表 3-5 所示，乔丹两个赛季的场均得分为 31.3，而科比达到了 33.4，明显高于乔丹，这就是反直觉的“辛普森悖论”。

表 3-5　乔丹和科比两个赛季得分合计

	球员	总得分	场次	场均得分
两个赛季合计	乔丹	5 014	160	31.3
	科比	3 310	99	33.4

“辛普森悖论”出现的关键因素是科比在前一个赛季仅出战 17 场，相比于 80 和 82，17 是个微不足道的小数字，因此，当两个赛季的得分相加后取均值时，科比前一个赛季的得分数据贡献很小，这就会导致悖论出现。“辛普森悖论”提醒我们，数据量相同或相近时才适合进行均值比较，否则会有失公允。

中位数与箱线图

中位数与箱线图是我们理解数据的另一种视角，接下来，我们用中位数和箱线图来分析科比 2008—2009 赛季常规赛的得分数据，看看它们与均值、直方图有什么不同。

中位数，顾名思义，就是处在中间位置上的数字。要找到中间位置，首先要对数据进行排序。表 3-6 是经过排序后的科比得分数据，从中找到排在中央的数据，便是中位数。如果有 81 个数据，第 41 个就是中位数，可是表 3-6 中有 82 个数字，我们需要取第 41 和第 42 个数的平均值作为中位数 M：

$$M=(37+37)/2=37$$

续表

16	23	28	34
16	23	28	35
16	23	28	36
17	23	28	36
18	23	29	36
18	24	29	36
18	24	29	37
19	25	29	37
19	25	30	38
19	26	30	39
20	26	30	39
20	26	30	40
20	26	31	41
20	26	31	49
21	27	32	61
21	27		

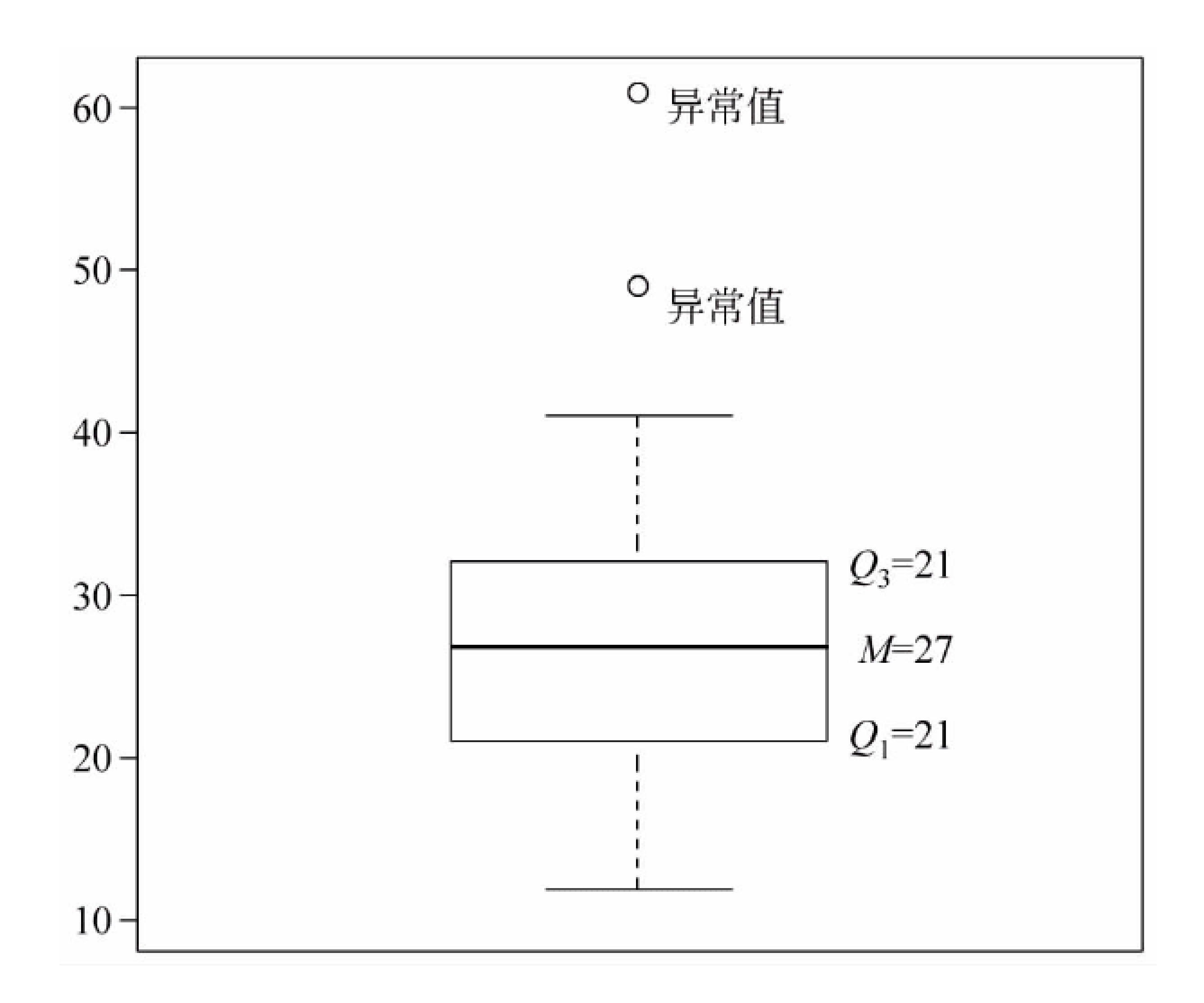


图 3-2　科比得分数据的箱线图

在箱线图中，区间的长度与数据的分散程度相关，比如，Min 到 Q_1 的长度是 11，Q_1 到 M 的长度是 6，M 到 Q_3 的长度是 5，Q_3 到 Max 的长度是 29，因此，M 到 Q_3 区间内，数据分布最集中，其次是 Q_1 到 M 的区间，数据分布最分散的区间是 Q_3 到 Max。

除了表征数据的分散程度，箱线图还可以帮助我们寻找疑似异常值。所谓疑似异常值是指过大或过小的数据，寻找的方法是：首先计算四分位数

差 IQR：

$$IQR = Q_3 - Q_1 = 32 - 21 = 11$$

然后找出小于 $Q_1 - 1.5\text{IQR}$ 和大于 $Q_3 + 1.5\text{IQR}$ 的数字，这些数字就是疑似异常值。

$$Q_1 - 1.5\text{IQR} = 21 - 1.5 \times 11 = 4.5$$

$$Q_3 + 1.5\text{IQR} = 32 + 1.5 \times 11 = 48.5$$

49 和 61 大于 48.5，所以是疑似异常值。在某些统计分析问题中，疑似异常值可能是误差数据甚至错误数据，可以通过上述方法找出并剔除这些数据，然后再绘制修正后的箱线图。对科比的得分数据来说，49 分和 61 分显然不是由误差或错误造成的，恰恰相反，这些“异常值”是“黑曼巴”贪婪本性的最佳诠释。

33 方差与标准差：致敬马刺

我是一个 NBA 的老球迷，回首将近 20 年的看球生涯，有一支球队让我不得不叹服，今天的他们仿佛从 20 年前穿越而来，“波波”还是那个“波波”，“石佛”还是那尊“石佛”，在其他球队经历大起大落的 20 年里，他们稳如泰山，从不动摇，你可以不喜欢他们的球风，但你必须尊重他们的坚守，他们是圣安东尼奥马刺队。

圣安东尼奥位于美国南部得克萨斯州，与达拉斯和休斯敦并称得州三大城市。1970 年，“得克萨斯橡木队”将主场移师圣安东尼奥，并更名为“马刺队”，马刺是指骑马者钉在鞋跟上的一种铁制的刺马针，是美国西部大开发的时代象征。初入 NBA 的 20 多年里，马刺队只能算是个不温不火的小角色，直到 1996—1997 赛季，这个赛季马刺队糟糕的战绩却意外的成就了他们未来 20 年的辉煌。由于 3 胜15 负的糟糕开局，球队总经理格雷格 · 波波维奇临危受命，担任球队主帅，随后，“凭借”糟糕的常规赛战绩，马刺队拿到了头号选秀权，蒂姆 · 邓肯空降圣城。自此以后，波波维奇与邓肯走上了 20 年的坚守之路。1998－1999 赛季，凭借邓肯与大卫 · 罗宾逊的内线“双塔”组合，马刺队夺得队史第一座冠军奖杯，2001 年和2002 年，“法国跑车”托尼 · 帕克和“阿根廷

妖刀”吉诺比利相继加盟球队，组成了日后马刺队的铁三角“GDP 组合”，马刺队在此后的近 20 年里再夺四次总冠军，他们永远是其他球队最不想遭遇的对手。

波波维奇教练是马刺队场下的灵魂，他秉承欧洲篮球的执教理念，进攻时强调快速转移球、球动人动，防守时强调持续逼抢和快速补位，再加上波波维奇的空军学院出身，马刺队俨然是一支训练有素的铁军，这支铁军的挂帅之人非邓肯莫属！邓肯，因球风沉稳、不苟言笑，江湖人称“石佛”，那几近失传的“45°打板投篮”最能体现邓肯朴实无华的球风，扎实的脚步移动、稳定的中距离投篮和遮天蔽日的封盖都是邓肯的标签。本赛季，绰号“圣安东尼奥养老院”的马刺队居然创造了队史常规赛胜场纪录，即将年满 40 岁的邓肯能否在职业生涯谢幕前再夺总冠军？我们拭目以待！

方差与标准差

马刺队的稳定令人惊叹，战绩可以说明一切，与同样在近 20 年夺得 5 次总冠军的湖人队相比，最能说明马刺队的稳定是多么可怕。表 3-7 是马刺队和湖人队自 1998 年以来的历年战绩，接下来，我们就用统计学的方法来说明，马刺队比湖人队更稳定。

表 3-7　马刺队和湖人队的历年常规赛战绩（1998—2015 年）

	马刺队				湖人队			
赛季	胜场	负场	胜率（%）	季后赛成绩	胜场	负场	胜率（%）	季后赛成绩
2014—2015	55	27	67.1	西部首轮	21	61	25.6	未进季后赛
2013—2014	62	20	75.6	总冠军	27	55	32.9	未进季后赛
2012—2013	58	24	70.7	总决赛	45	37	54.9	西部首轮
2011—2012	50	16	75.8	西区决赛	41	25	62.1	西部半决赛
2010—2011	61	21	74.4	西部首轮	57	25	69.5	西部半决赛
2009—2010	50	32	61.0	西区半决赛	57	25	69.5	总冠军
2008—2009	54	28	65.9	西部首轮	65	17	79.3	总冠军
2007—2008	56	26	68.3	西区决赛	57	25	69.5	总决赛
2006—2007	58	24	70.7	总冠军	42	40	51.2	西部首轮
2005—2006	63	19	76.8	西区半决赛	45	37	54.9	西部首轮

续表

	马 刺 队				湖 人 队			
赛季	胜场	负场	胜率（%）	季后赛成绩	胜场	负场	胜率（%）	季后赛成绩
2004－2005	59	23	72.0	总冠军	34	48	41.5	未进季后赛
2003－2004	57	25	69.5	西区半决赛	56	26	68.3	总决赛
2002－2003	60	22	73.2	总冠军	50	32	61.0	西部半决赛
2001－2002	58	24	70.7	西区半决赛	58	24	70.7	总冠军
2000－2001	58	24	70.7	西区决赛	56	26	68.3	总冠军
1999－2000	53	29	64.6	西部首轮	67	15	81.7	总冠军
1998－1999	37	13	74.0	总冠军	31	19	62.0	西部半决赛
平均值	57.5	24.5	70.1		49.1	32.9	59.9	
标准差	3.48	3.48	4.24		13.43	13.43	16.38	

在表 3-7 的数据中，1998—1999 赛季和 2011－2012 赛季是两个缩水的赛季，比赛场次较少，为了避免掉入“辛普森悖论”的陷阱，我们将这两行数据排除在外（以深色标记），其余赛季的总场次都是 82 场，因此，胜率可以进行对比和加减运算。

我们首先计算两支球队的平均胜率，X 和 Y 分别代表马刺队和湖人队。

马刺队平均胜率 $\overline{X} = (67.1\% + 75.6\% + \cdots + 64.6\%)/15 = 70.1\%$

湖人队平均胜率 $\overline{Y} = (25.6\% + 32.9\% + \cdots + 81.7\%)/15 = 59.9\%$

读者还可以试着画一画两队胜率的直方图和箱线图，不论怎样，我们都必须承认，马刺队的成绩总体上优于湖人队。接下来，我们算一算马刺队到底有多稳定。

方差和标准差是统计学中用于描述数据发散程度的统计量，假设有数据 $X_1=1$ 和 $X_2=3$，均值为 $\overline{X}=2$，那么，方差为：

$$Var(X) = [(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2]/2 = [(1-2)^2 + (3-2)^2]/2 = 1$$

标准差为

$$\sigma(X) = \sqrt{Var(X)} = 1$$

如果是 n 个数据 $X_1 \sim X_n$，均值为 $\overline{X}$，则方差为

$$Var(X) = [(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]/n$$

标准差为

$$\sigma(X) = \sqrt{[(X_1 - \overline{X})2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]/n}$$

这正是均方根值的用处。

利用上面的公式，可以计算出马刺队和湖人队的胜率方差和标准差：

马刺队胜率方差 $Var(X)=[(67.1\%-70.1\%)^2+\cdots+(64.6\%-70.1\%)^2]/15=0.18\%$；

马刺队胜率标准差 $\sigma(X)=4.24\%$；

湖人队胜率方差 $Var(Y)=[(25.6\%-59.9\%)^2+\cdots+(81.7\%-59.9\%)^2]/15=2.68\%$；

湖人队胜率标准差 $\sigma(Y)=16.38\%$。

标准差与均值有相同的单位，是可以比较的，因此，综合来看：

马刺队的平均胜率是 70.1%，标准差是 4.24%；

湖人队的平均胜率是 59.9%，标准差是 16.38%。

在统计学中，标准差越小代表数据的分布越集中于均值附近。马刺队的胜率标准差远小于湖人队，意味着他们的胜率更集中的分布在均值周围，这便是马刺队令全联盟生畏的稳定性。

3.4　均值与方差估计：近射与狙击

枪是很多儿童的最爱，手握玩具枪，扮演警察叔叔，“biu”的一枪击毙坏蛋，是儿童永远玩儿不腻的游戏。我国禁止枪支买卖，因此大多数人都没有机会摸枪，最多是在大学的军训课上匆匆扣几次扳机了事。我曾在大学里选修射击课，练习过手枪和步枪射击，因此对射击有了更多的体验。

射击从目标距离上大致分为两类：一类是近距离射击，一般使用手枪；另一类是远距离狙击，一般使用步枪或狙击枪。不管是哪一类射击，最要紧的就是一个字——准。要射得准，先要瞄得准。瞄准有方法和经验可循，近距离的手枪射击，只要保持手型端正，按照“三点一线”的要求，把缺口、准星和目标点连成一条线便可以；100 米的步枪射击或狙击枪射击，仅靠“三点一线”是不够的，瞄准时，我们不能把靶心 10 环设为目标点，而是要把下 8 环甚至下 7 环设为目标点，这样才能射中靶心，这是对教科书的合理校正。

在统计学中，常要通过样本来估计总体的均值和方差，这两种估计也都讲

究一个“准”字，统计学中称之为“无偏”，二者的估计方法并不相同，与近距离射击和远距离狙击有异曲同工之妙。

表 3-8 所示是科比 82 场常规赛得分数据的样本和总体，我们以表中数据为例，说明样本对总体的均值和方差估计。

表 3-8　科比得分数据的总体和样本

样本数据	总体数据			
11	10	21	27	32
16	11	22	28	32
18	11	22	28	33
19	12	22	28	33
20	14	23	28	33
22	16	23	28	34
23	16	23	28	35
23	16	23	28	36
25	17	23	28	36
26	18	23	29	36
28	18	24	29	36
28	18	24	29	37
29	19	25	29	37
30	19	25	30	38
31	19	26	30	39
33	20	26	30	39
35	20	26	30	40
36	20	26	31	41
39	20	26	31	49
41	21	27	32	61
	21	27		

样本共有 20 个数据，记为 $X_1, X_2, \cdots, X_{20}$，总体共有 82 个数据，记为 $Y_1, Y_2, \cdots, Y_{82}$。

首先来看总体均值估计。

样本是一个“迷你版”的总体，只要采样足够随机，样本应与总体有相似的分布特征，因此，我们可以用样本的均值来估计总体的均值。

在本例中，样本均值为

$$\overline{X} = (X_1 + X_2 + \cdots + X_{20}) / 20 = 26.65$$

总体均值记为 μ,其估计值记为 $\hat{\mu}$,因此有

$$\hat{\mu} = \overline{X} = 26.65$$

实际上,总体均值为

$$\mu = (Y_1 + Y_2 + \cdots + Y_{82}) / 82 = 26.84$$

可见 $\hat{\mu}$ 和 μ 很接近。用样本均值估计总体均值与手枪近射类似,瞄哪儿打哪儿。

再来看总体方差估计。

前面我们提到,样本应与总体有相似的分布特征,因此我们自然认为,样本的方差也应该代表总体的方差。

在本例中,样本方差为

$$Var(X) = [(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_{20} - \overline{X}^2)] / 20 = 59.12$$

总体方差记为 σ^2,其估计值记为 $\hat{\sigma}^2$,按照我们此前的推理,

$$\hat{\sigma}^2 = Var(X) = 59.12$$

实际上,总体方差为

$$\sigma^2 = [(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + \cdots + (Y_{82} - \overline{Y})^2] / 82 = 72.23$$

可见,$\hat{\sigma}^2$ 比 σ^2 要小一点。这并不是特例,而是普遍现象,样本的方差往往比总体方差要小一点。可是,“一点”是多少呢?这很难说得清,但统计学家们还是找到了弥补这“一点”的方法:把样本方差计算公式中的分母变由 n 变为 $n-1$,使样本方差变大“一点”。

经过修正后的方差称为无偏方差,记为 S^2,例子中样本的无偏方差为

$$S^2 = [(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_{20} - \overline{X})^2] / (20 - 1) = 62.24$$

相比于修正之前,无偏方差更接近总体方差。用 S^2 来估计总体方差与远距离狙击类似,都要做出适量的校正。

看到这里,读者很可能会有这样的疑问:为什么不是 $n-2$,$n-3$?

单从这个例子来看,取 $n-2$ 或 $n-3$ 都比 $n-1$ 的估计效果更好,但这只是一个特例。采用 S^2 来估计总体方差并非经验式修正,是有数学理论依据的,感兴趣的读者可以参考概率统计的专业书籍。

这里需要说明的是,上一节中,我们并没有把胜率数据看作样本,因此没有使用修正后的方差公式。对一组独立数据来说,方差就是 Var,不是 S^2。如

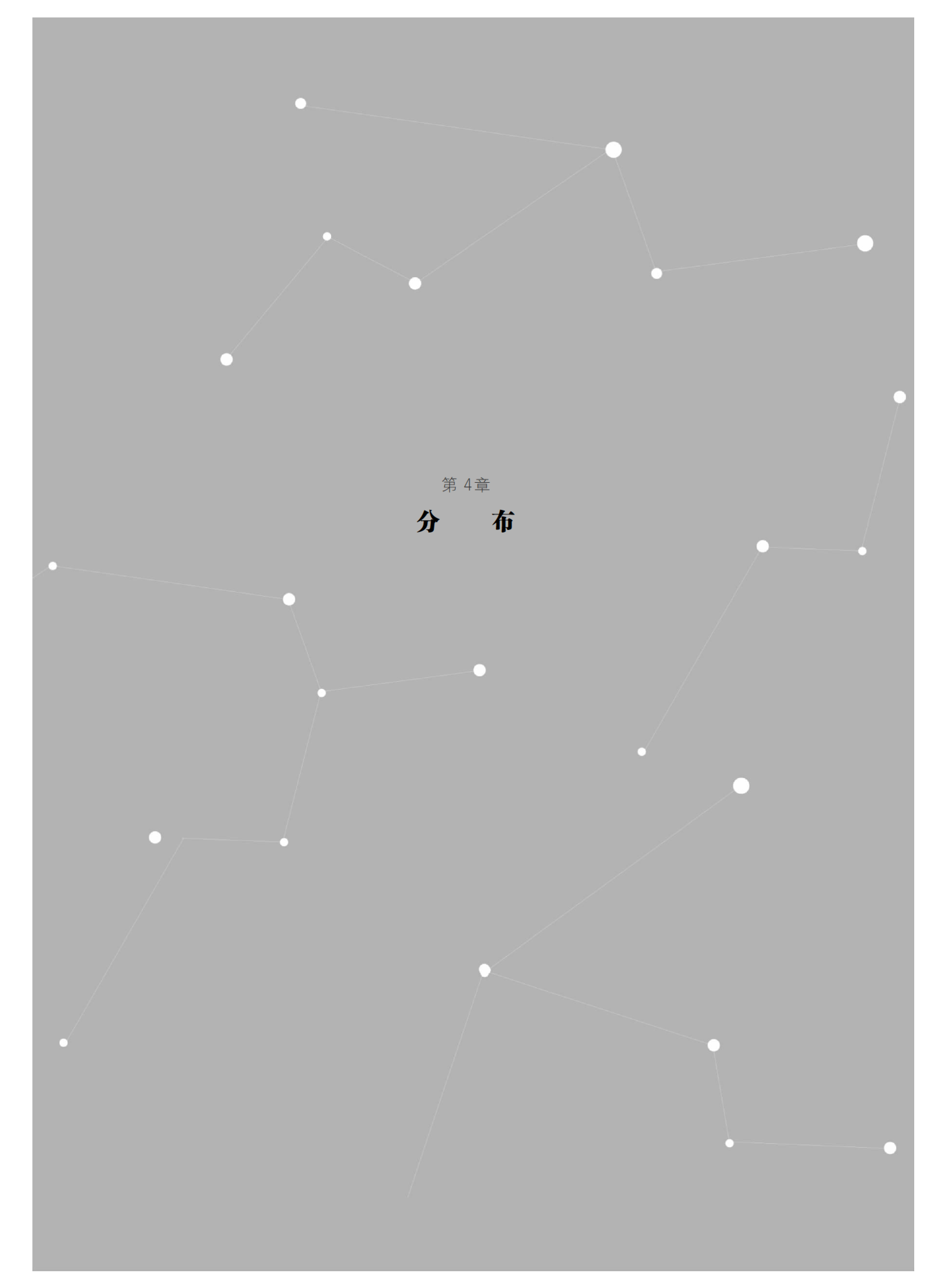
果你把数据看作总体的样本，方差就是 S^2，这两者的区别读者一定要留心。

最后，总结一下均值和方差估计。

设 $X_1, X_2, \cdots, X_n$ 是来自总体的样本，那么，总体的均值和方差的无偏估计分别是

$$\hat{\mu} = (X_1 + X_2 + \cdots + X_n)/n$$

$$\hat{\sigma}^2 = [(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]/(n-1)$$



第 4章

分　　布

导语：“小九九”是乘法运算的本源，千变万化的乘法运算都是从“小九九”演化而来的。概率分布就像概率统计的“小九九”，它可以帮助我们解决很多常见的概率统计问题，既简洁又高效。

4.1 分布：统计学的“小九九”

不管喜不喜欢数学课，你一定记得“小九九”，你一定知道“一一得一，一二得二”和“九九八十一”。“小九九”是学习乘法的第一课，也是最重要的乘法口诀。常言道：“万变不离其宗”，“小九九”便是乘法之“宗”，千变万化的乘法运算都是从“小九九”演化而来的。

统计学也有自己的“小九九”，它不是一个口诀，而是从很多典型概率问题中总结出的经验，我们称为概率分布，简称分布。

分布是随机变量的取值与其对应概率的关系。例如，抛硬币试验中，设反面为 0，正面为 1，随机变量 X 为抛出硬币的数值，X 的分布如表 4-1 所示。又如，掷骰子试验中，随机变量 Y 为掷出的点数，Y 的分布如表 4-2 所示。表 4-1

和表 4-2 就是随机变量的分布，利用分布，可以计算出随机变量的期望和方差。

表 4-1　抛硬币试验中随机变量 X 的分布

X 取值	概率
0	1/2
1	1/2

表 4-2　掷骰子试验中随机变量 Y 的分布

Y 取值	概率
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

统计学家可不想一个个地列出随机变量的分布，他们要对随机变量归类，计算出同一类随机变量的分布、期望和方差。对上面的两个例子来说，抛硬币和掷骰子都属于等概率分布，即随机变量每个取值的概率都相等。如果我们知道了等概率分布的计算公式，就不需要列表格了，直接做个“伸手党”，套用公式就可以了，这就是统计学家研究分布的原因，也是我们学习分布的原因。

在开始学习分布之前，再次提醒读者，随机变量分离散和连续两类，分别对应离散分布和连续分布。虽然本书前面的内容都是有关离散随机变量的，但是连续随机变量和连续分布在概率统计中也占有重要地位。因此，常用的离散分布和连续分布都是需要我们学习的。

下面，我们就一起来学习常用的几种概率分布。

4.2　等概率分布：硬币的两面

抛硬币是概率论中最常见的随机试验，不仅因为硬币很常见，也因为抛硬币试验中，随机变量的分布是最简单的等概率分布。

等概率分布，顾名思义，就是随机变量每一个取值的出现概率都相等。在概率论的发展初期，等概率分布是主要研究对象，后人也把与抛硬币、掷骰子相似的随机试验称为“古典概型”。下面，我们使用“从特殊到一般”的归纳思想来学习等概率分布。

以抛硬币为例，反面记为 0，正面记为 1，随机变量 X 为抛硬币一次的得分，那么，X 的分布可以写为

$$P(X=k)=\begin{cases}\frac{1}{2}, & k=0\\ \frac{1}{2}, & k=1\end{cases}$$

X 的期望是

$$E(X)=0\times P(X=0)+1\times P(X=1)=1/2$$

X 的方差是

$$\begin{aligned}Var(X)&=P(X=0)\times[0-E(X)]^2+P(X=1)\times[1-E(X)]^2\\ &=\{[0-E(X)]^2+[1-E(X)]^2\}/2\end{aligned}$$

再以掷骰子为例，随机变量 Y 为掷一个骰子的点数，那么，Y 的分布可以写为

$$P(Y=k)=\begin{cases}\frac{1}{6}, & k=1\\ \frac{1}{6}, & k=2\\ \frac{1}{6}, & k=3\\ \frac{1}{6}, & k=4\\ \frac{1}{6}, & k=5\\ \frac{1}{6}, & k=6\end{cases}$$

Y 的期望是

$$\begin{aligned}E(Y)&=1\times P(Y=1)+2\times P(Y=2)+\cdots+6\times P(Y=6)\\ &=(1+2+\cdots+6)/6=3.5\end{aligned}$$

Y 的方差是

$$Var(Y) = P(Y=1) \times [1-E(Y)]^2 + P(Y=2) \times [2-E(Y)]^2 + \cdots + P(Y=6) \times [6-E(Y)]^2$$
$$= \{[1-E(Y)]^2 + [2-E(Y)]^2 + \cdots + [6-E(Y)]^2\}/6$$

我们仔细观察上面的分布和期望、方差计算公式,可以从这些个例中归纳出等概率分布的通用表达。

随机变量 X 有 n 个取值 $a_1, a_2, \cdots, a_n$,每个取值出现的概率相等,那么,随机变量 X 的分布可以记为

$$P(X = a_k) = 1/n, \quad k = 1, 2, \cdots, n$$
$$E(X) = (a_1 + a_2 + \cdots + a_n)/n = \Sigma a_k/n$$
$$Var(X) = \{[a_1 - E(X)]^2 + [a_2 - E(X)]^2 + \cdots + [a_n - E(X)]^2\}/n$$
$$= \sum [a_k - E(X)]^2/n$$

(注:$\sum$ 是求和符号,表示对 k 的不同取值求和。)

上面的三个公式便是等概率分布的分布、期望和方差的计算公式,再次遇到等概率分布的问题时,我们可以直接使用这些公式来计算分布、期望和方差。

等概率的陷阱

等概率分布是最简单的概率分布,看似简单的表象下,却隐藏着思维陷阱。

此前抛硬币的例子只提到了抛掷硬币一次,如果抛掷多次会怎样呢?下面,请用最快的速度回答下面的问题:

抛掷硬币 10 次,“正正正正正正正正正正”与“正正反正反反正反正反”哪一个更可能出现?

你的直觉很可能是:后者更可能出现。而正确答案是:两种情况出现的可能性是一样的,都是$(1/2)^{10}$。其实,大多数人都会有这样的错觉:十次全是正面,这太特殊了,不太可能出现。这种错觉很可能导致你的错判——认为后一种情况更可能出现,因为它看起来更“正常”。这里必须提醒读者,假如我们要严谨地思考一个与概率有关的问题,千万不要相信感觉,最靠谱的方法是动

笔算一算。

估计上面的陷阱并没有把你骗进去，下面我们来看一个逻辑悖论——钱包悖论。

假设你的面前有两个钱包，其中一个钱包里的钱是另一个的两倍。你随机选择一个钱包，打开它，发现里边装着 100 元，请问，你是决定留下这个钱包还是丢下它选择另一个钱包呢？

如果仅凭直觉，大多数人会为了得到 200 元选择另一个钱包。巧合的是，这一次概率论和我们的直觉不谋而合。我们不知道另一个钱包里是 200 元还是 50 元，所以，这两种情况出现的可能性各为 1/2，所以，换钱包的收益期望是：

$$(-50)\times 1/2+100\times 1/2=25(\text{元})$$

的确是正数！赶紧换钱包吧！

等等，先别着急数钱，回想一下这个游戏，一个非常有趣的局面出现了。上述逻辑可以简化为：不论第一个钱包里装了多少钱，你都会选择另一个钱包。言外之意，你根本不需要打开第一个钱包，只要随机选一个，然后换第二个就可以了，可是，这跟直接选第二个难道不一样吗？更让人抓狂的是，一旦你打开了第二个钱包，这个钱包就变成了你随机选的第一个钱包了，于是，你决定换回第一个钱包。

莫非是打开的方式不对？

第一次的选择到底有没有意义？

如果我永远不打开钱包，岂不是要永远换下去，而且越换赚的钱越多！？

别再纠结这些问题了，其实我们刚开始便犯了一个致命的错误——认为未知的情况都是等概率出现的。题目里说，一个钱包里的钱是另一个的两倍，可是，这并非意味着 200 元和 50 元出现的概率相同，我们“不知道”它们出现的概率是多少，并不代表它们出现的概率相同，事实上，这里根本就不存在概率，不能用概率来解释！

最后，我们换一种方式描述这个悖论：你的面前有两个钱包，一个钱包里有 A 元，另一个有 $2A$ 元，你随机选择一个，打开，然后选择另一个钱包。这时，你得到 A 元和失去 A 元的概率是相等的。这才是两个钱包正确的打开方式！

4.3 几何分布：一次就好

一次就好我带你去看天荒地老
在阳光灿烂的日子里开怀大笑
在自由自在的空气里吵吵闹闹
你可知道我唯一的想要

杨宗纬《一次就好》

2015 年开心麻花出品了电影处女作《夏洛特烦恼》,《一次就好》是夏洛追求校花时唱起的情歌。“一次就好”让人既温暖又唏嘘,可是,追求心爱的人从来都不会一帆风顺,只有不断尝试,越挫越勇,才能收获爱情。

有一个很特别的分布,叫作几何分布,这个分布告诉人们,什么时候才能实现第一次。

仍以抛硬币为例,已知出现正反两面的概率各为 1/2,在反复抛掷的过程中,我们设定随机变量 X 表示第一次出现反面时抛掷硬币的次数,我们列出 X 的概率分布,如表 4-3 所示。

表 4-3　第一次出现反面时抛掷硬币的次数 X 的分布

X	$P(X)$
1	1/2
2	(1/2)×(1/2)=1/4
3	(1/2)×(1/2)×(1/2)=1/8
4	(1/2)×(1/2)×(1/2)×(1/2)=1/16
…	…

用数学公式来表达为

$$P(X=k)=(1/2)^{k-1}\times(1/2),\quad k=1,2,3,\cdots$$

式中,$(1/2)^{k-1}$ 表示前 $k-1$ 次都是正面,乘号后边的 1/2 表示第 k 次是反面。

这个例子有些特殊,因为正面和反面出现的概率相同,如果不相同会怎样呢?我们以骰子游戏为例。

已知骰子有六种点数，每个点数出现的概率都是 1/6，反复抛掷骰子，设定随机变量 Y 表示第一次出现六点时抛掷骰子的次数，我们列出 Y 的概率分布，如表 4-4 所示。

表 4-4　第一次出现六点时抛掷骰子的次数 Y 的分布

Y	P(Y)
1	1/6
2	(5/6)×(1/6)=5/36
3	(5/6)×(5/6)×(1/6)=25/216
4	(5/6)×(5/6)×(5/6)×(1/6)=125/1 296
…	…

用数学公式来表达为

$$P(Y=k)=(5/6)^{k-1}\times(1/6),\quad k=1,2,3,\cdots$$

式中，$(5/6)^{k-1}$ 表示前 $k-1$ 次都不是六点，1/6 表示第 k 次是六点。

透过两个例子，我们可以归纳出几何分布的通用表达。

设随机试验有且只有两种结果 A 和 B，A 出现的概率是 p，B 出现的概率是 $1-p$，反复进行该随机试验，随机试验之间彼此独立，随机变量 X 表示 A 第一次出现时随机试验进行的次数，此时我们称随机变量 X 服从几何分布：

$$P(X=k)=(1-p)^{k-1}\cdot p,\quad k=1,2,3,\cdots$$

图 4-1 是几何分布的概率分布图，从图中可以明显地看出，虽然 X 的取值有无穷多个，但是 $X=1$ 的概率是最大的，也就是说，1 次成功的可能性最大。

几何分布是一个无限可列的概率分布，要计算它的期望和方差需要使用一些数列求和的计算技巧，我们不细究这些计算技巧，直接给出几何分布的期望和方差。

$$E(X)=1/p$$

$$Var(X)=(1-p)/p^2$$

几何分布的期望与我们的直觉不谋而合。比如，硬币出现反面的概率是 1/2，那么平均意义上需要抛 2 次才会出现反面；骰子的六点出现的概率是 1/6，那么平均意义上需要掷 6 次才能出现六点；中一次彩票大奖的概率是百万分之一，那么平均意义上需要买一百万次才能中一次大奖。

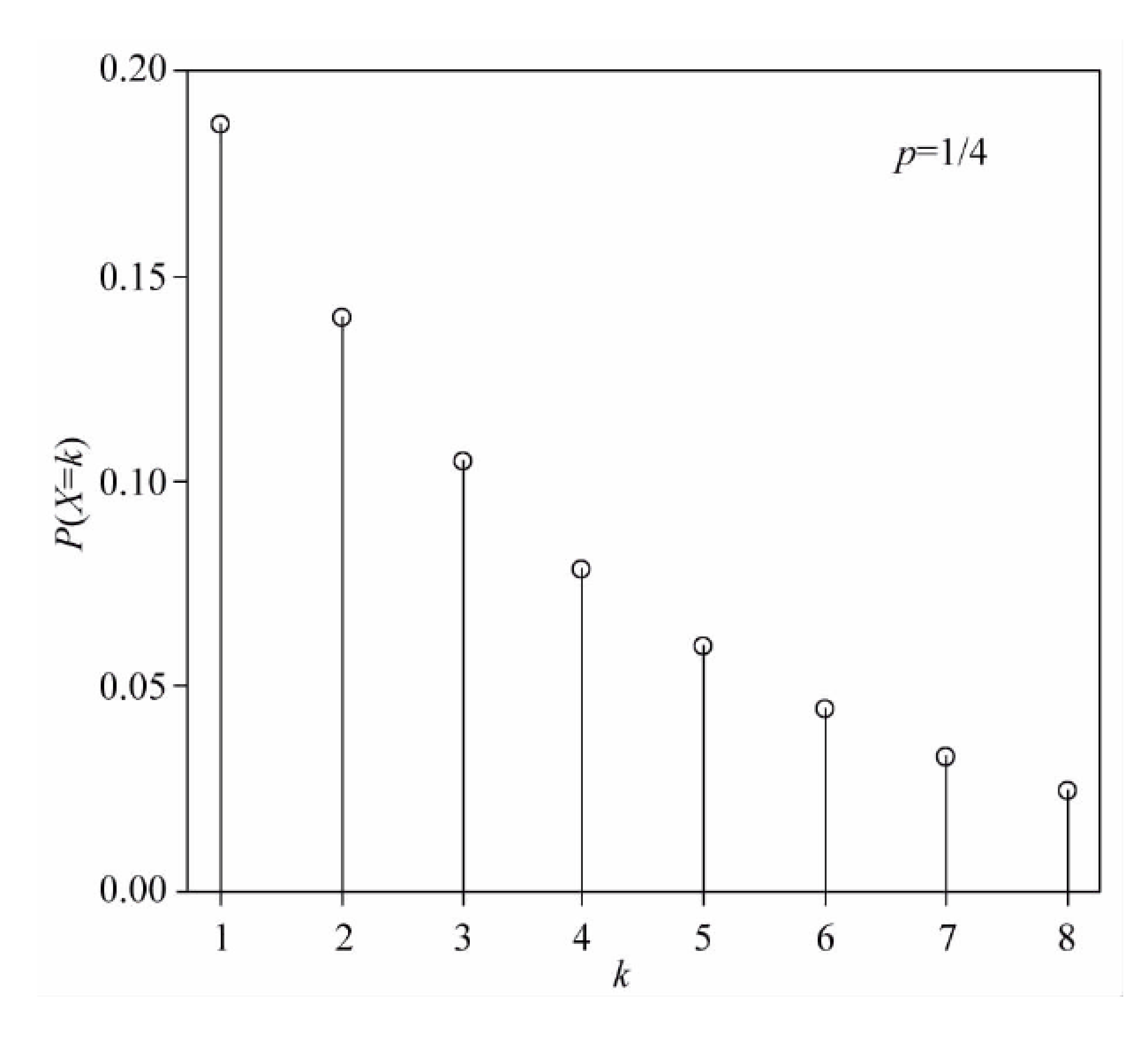


图 4-1　几何分布

几何分布只适用于反复进行的独立试验，这一点很容易被人们忽视，我们用两个例子来说明。

例 1：选手 A 参加"一站到底"的选拔考试，题目分三类，历史类、体育类和文学类，每一轮答题，A 要从三类题目的混合题库中随机抽取一道题作答。假设 A 只擅长历史类问题，那么，A 答对第一道题平均需要多少轮？

例 2：选手 A 参加"一站到底"的选拔考试，题库只有三个问题，分别属于历史类、体育类和文学类，每一轮答题，A 要从三个问题中随机抽取一个作答，作答后该题随即作废。假设 A 只擅长历史类问题，那么，A 答对第一道题平均需要多少轮？

这两个例子类似于抽样中的重复抽样和不重复抽样。例 1 属于重复抽样，A 每一轮答题彼此独立，而且答对的概率相同，都是 1/3，因此，例 1 是典型的几何分布，期望是 1/(1/3)＝3，所以例 1 的答案是 3 轮。在例 2 中，如果 A 第一轮没答对，第二轮答对的概率就会变为 1/2，如果进入第三轮，他答对的概率更是 100%，每轮答题的结果会改变后面轮次的概率，因此各轮之间不是互相独立的，所以例 2 不能用几何分布来解释。我们简单计算一下便会发现，例 2 中 A 在第一、二、三轮首次答对的概率都是 1/3，因此，他首次答对问题所需的平均轮次是(1/3)×(1＋2＋3)＝2，即 A 平均只需要两轮就可以答对一个

问题。两个相对比,不重复抽样的规则更有利于 A。

4.4　二项分布:反复掷骰子

拿来一副扑克牌,抽出大小王,剩下 52 张牌,这 52 张牌分属黑桃、红桃、梅花、方块四种花色,把这 52 张牌随机的发给 4 位玩家,每人 13 张牌。定义花色分布为四种花色的牌数组合,并且与花色无关,例如,4－4－3－2 是一种花色分布,5－5－3－0 是另一种花色分布。请问:这 13 张牌最可能的花色分布是怎样的?

是看似最平均的 4－3－3－3? 还是其他花色分布? 二项分布将会告诉我们答案。

二项分布来源于伯努利试验,所谓伯努利试验就是只有两种可能结果的随机试验,比如抛硬币。当一个伯努利试验独立地重复进行 n 次时,几何分布只能告诉我们第一次何时发生,二项分布则可以告诉我们各种可能的结果发生的概率。接下来,我们就从几何分布出发一起来认识二项分布。

几何分布告诉我们,掷骰子时,平均意义上需要 6 次才会第 1 次出现六点。一个赌场老板知道了这个结论,信心满满地开设了一个赌局:掷骰子 5 次,如果六点一次都没出现,庄家赢;否则,庄家输。他的想法是,既然平均要 6 次才会第 1 次出现六点,那么掷 5 次不出现六点的概率肯定比出现六点的概率要高,庄家稳赚不赔。这个想法听起来很靠谱,到底对不对,我们来算一算。

骰子每次出现六点的概率依然是 1/ 6,不出现六点的概率是 5/ 6,我们要计算掷骰子 5 次至少出现 1 次六点的概率。请读者们忘记逆向思维,要真正认识二项分布,需要从正面来思考。至少出现 1 次,可以分为出现 1 次、2 次、3 次、4 次和5 次共 5 种情况,“出现 1 次”又可以分为仅第 1 次出现、仅第 2 次出现、……、仅第 5 次出现共 5 种情况,如此这般,穷举所有情况,一定可以计算出结果。除了穷举法,我们还可以利用一个数学工具,使计算变得简单,这个数学工具就是排列组合。

排列组合

排列组合是用来解决诸如“从牌堆里取出若干张牌，有多少种可能的牌型”这类问题的数学公式，分为排列公式和组合公式两类，排列是有序的，组合是无序的。我们以斯诺克台球为例，来学习排列组合的基础知识。

斯诺克台球比赛开球时，除了白色母球外，球桌上有 15 颗红球和 6 颗彩色球，红球彼此相同，分值为 1 分，彩色球各不相同，按照分值由低到高分别为黄色球（2 分）、绿色球（3 分）、棕色球（4 分）、蓝色球（5 分）、粉色球（6 分）和黑色球（7 分）。

组合问题：将两颗红球随机放进 6 个球袋中的 2 个，有几种放置方法？

因为红球彼此相同，我们使用组合公式来计算，计算方法是：

$$\text{放置种类} = C_6^2 = 6!/(2!\times 4!) = 15\ \text{种}$$

[注：!是阶乘符号，对任意整数 k，k!读作“k 的阶乘”，表示 $k\times(k-1)\times\cdots\times 2\times 1$]

排列问题：将蓝色球和粉色球随机放进 6 个球袋中的 2 个，有几种放置方法？

蓝色球和粉色球彼此不同，同样是放置在 1 号球袋和 2 号球袋中，有两种放置方法，而两颗红球只有一种放置方法，如图 4-2 所示。因此，我们使用排列公式A_6^2来计算，计算方法是

$$\text{放置种类} = A_6^2 = 6!/4! = 30\ \text{种}$$



图 4-2　排列与组合的区别

将上述公式进行归纳，便可以得到排列组合的通用表达。

将 k 个相同的球放进 n 个球袋中的 k 个，是组合问题，共有 $C_n^k=n!/[k!\cdot(n-k)!]$ 种放置方法。

将 k 个互不相同的球放进 n 个球袋中的 k 个，是排列问题，共有 $A_n^k=n!/(n-k)!$ 种放置方法。

二项分布

回到掷5次骰子的问题中，我们要分别计算六点出现1次、2次、3次、4次和5次的概率。

首先计算六点只出现1次的概率，我们可以把所有可能的情况一一列举出来，这些情况各自出现的概率都是 $(5/6)^4\times(1/6)$，可是，有多少种可能的情况呢？每次掷的骰子是相同的，所以，这是一个组合问题，一共有 C_5^1 种可能的情况，将所有可能情况的概率相加，便得到了六点出现1次的概率：

$$P(\text{六点出现 1 次})=C_5^1\times(5/6)^4\times(1/6)=0.4019$$

同理，可以计算出其他情况的概率：

$P(\text{六点出现 2 次})=C_5^2\times(5/6)^3\times(1/6)^2=0.1607$；

$P(\text{六点出现 3 次})=C_5^3\times(5/6)^2\times(1/6)^3=0.0321$；

$P(\text{六点出现 4 次})=C_5^4\times(5/6)^1\times(1/6)^4=0.0032$；

$P(\text{六点出现 5 次})=C_5^5\times(5/6)^0\times(1/6)^5=0.0001$。

将这五个概率相加便得到了六点至少出现1次的概率：

$$P(\text{六点至少出现 1 次})=0.5981$$

这个概率值大于0.5，说明掷5次骰子至少出现1次六点的概率更大，庄家不可能通过这个赌局赚到钱！其实，即使把规则改为掷4次骰子，至少出现1次六点的概率也有0.5177，还是大于0.5，庄家依然是输家。

把上面的计算方法进行归纳，便可以得到二项分布的通用表达。

设伯努利试验有两种可能结果 A 和 B，事件 A 发生的概率是 p，事件 B 发生的概率是 $1-p$，独立地重复进行 n 次该试验，设随机变量 X 表示事件 A 发生的次数，我们称随机变量 X 服从参数为 n,p 的二项分布，记为 $X\sim b(n,p)$，并且

$$P(X=k)=C_n^k\cdot(1-p)^{n-k}\cdot p^k$$

二项分布的期望和方差分别是：

$$E(X)=np$$

$$Var(X)=np(1-p)$$

在“大数定理”一节中，我们曾经提到过，反复抛掷硬币，正反面出现次数相等的概率会随着抛掷次数的增加越来越小，现在，我们就来计算一下正反面出现次数相等的概率。

抛掷硬币 10 次，出现 5 次正面 5 次反面的概率是：

$$P=C_{10}^5\times(1/2)^5\times(1/2)^5=0.246\ 1$$

抛掷硬币 100 次，出现 50 次正面 50 次反面的概率是：

$$P=C_{100}^{50}\times(1/2)^{50}\times(1/2)^{50}=0.079\ 6$$

抛掷硬币 1000 次，出现 500 次正面 500 次反面的概率是：

$$P=C_{1000}^{500}\times(1/2)^{500}\times(1/2)^{500}=0.025\ 2$$

对比这三个概率值可以发现，抛硬币的次数越多，正反两面出现次数相同的概率越小。

二项分布是一个十分独特的分布，我们从它的分布图中可以看出些端倪。图 4-3 给出的是 $b(10,1/2)$、$b(10,1/3)$、$b(10,1/5)$和 $b(10,1/10)$的概率分布图，我们观察四张图中的最高点：当 $p=1/2$ 时，概率最高点出现在 $X=5$ 的位置，概率分布关于最高点左右对称，当 $p=1/3$、$1/5$ 和 $1/10$ 时，概率分布不再对称，最高点的位置分别出现在 $X=3$、$X=2$ 和 $X=1$，是不确定的。从这组分布图中可以看出，二项分布并没有固定的规律可循，只有画出概率分布图才能找到最高点，即概率的最大值。

本节的最后，我们要回答开头提出的扑克牌问题了。13 张牌，4 种花色，最可能的花色分布是哪一种呢？是 4—3—3—3 吗？这个问题虽然不能直接用二项分布来计算，但是也具有二项分布相似的特征——“最平均的情况”未必是概率最大的。表 4-5 是列出了部分花色分布的概率，概率最高的花色分布果真不是 4—3—3—3，而是 4—4—3—2！

请读者们记住这个反直觉的案例，它将始终提醒着你：平均的未必是最可能发生的！

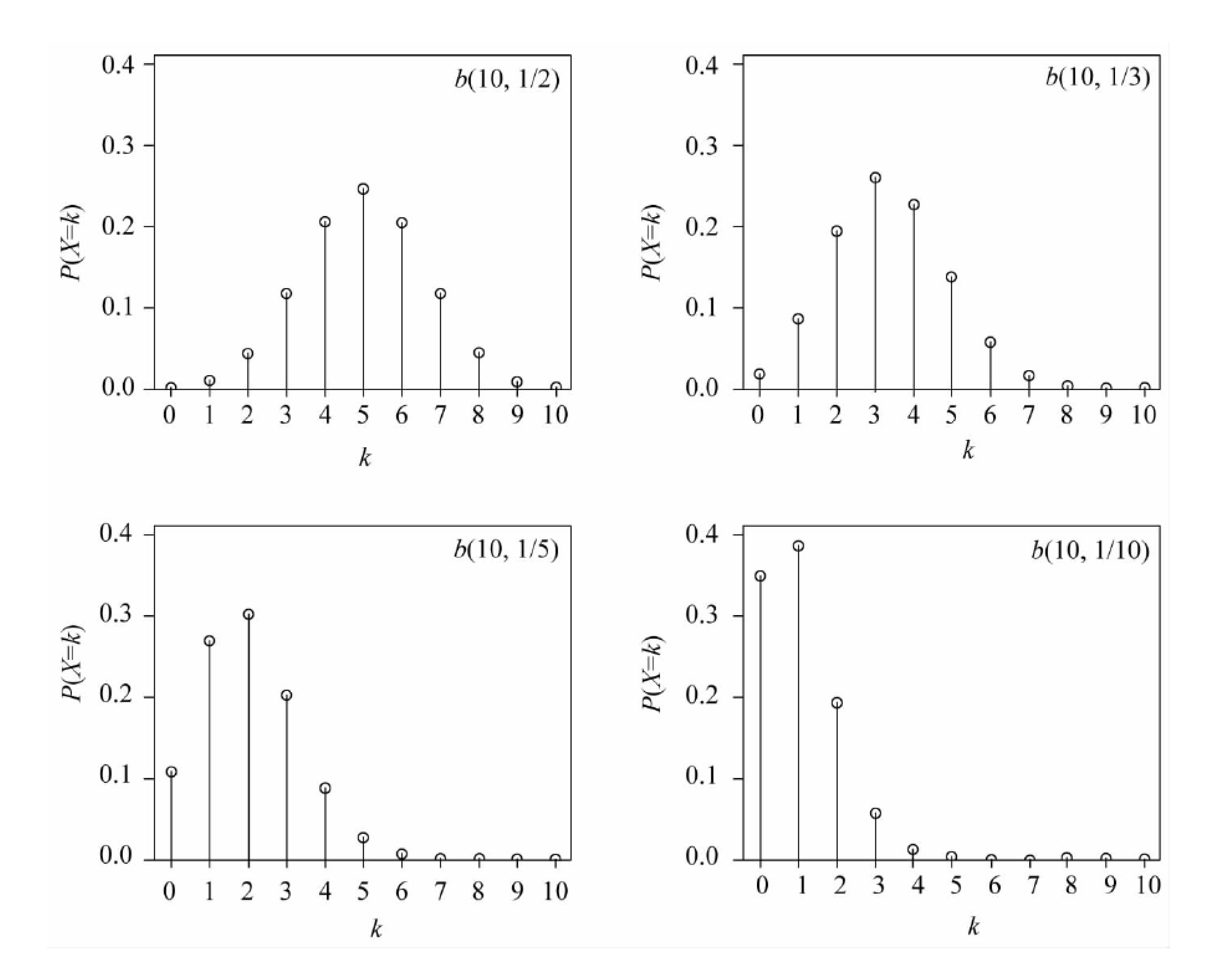


图 4-3　四个二项分布的概率分布图

表 4-5　花色分布的概率

花色分布	概率(%)(由大到小排序)
4-4-3-2	21.6
5-3-3-2	15.5
5-4-3-1	12.9
5-4-2-2	10.6
4-3-3-3	10.5
6-3-2-2	5.6
…	…

4.5　泊松分布：神奇的 e

如果你每天走在路上，被鸟粪砸中的概率刚好是 1/365，你一年里一次都没被砸中的概率是多少？

如果你是一个守株待兔的猎人，每天有兔子撞到树上的概率是 1/1 000，3 年里你一只兔子都没逮到的概率是多少？

如果飞机失事的概率是百万分之一，你坐一百万次飞机还没遇到事故的概率是多少？

这些问题的答案全部都是 37%。

清华园里有很多鸟儿，平日里走在林荫路上，真的可能被鸟粪砸中，我很幸运，大学四年一次都没被砸中，反而是我的一个外校同学，第一次来清华游玩就被鸟粪砸个正着。这就是让人无法预测的小概率事件，这些事件的确发生过，未来也有可能再次发生，可是谁也不知道它什么时候发生，它像幽灵一般神秘莫测。即便如此，统计学家们还是找到了其中的规律，我们先从 37%这个神奇的数字说起。

神奇的常数 e

37%，这个数字对大多数人来说很陌生，或许只有数学家才会知道，这个数字正是 1/e 的值。e 是自然对数底，是个无限不循环小数，数值为 2.718 2…。提起数学中的常数，大多数人会首先想到 π，其实，自然对数底 e 也是数学世界中十分重要的常数。下面我们就通过一个复利的小故事告诉你 e 的由来。

有一天，一个生意人急着用钱，便向一个财主借钱。财主见生意人十分着急，便趁机抬高利息，他开出的条件是，生意人每借 1 两银子，就要在一年后还 2 两银子，利率高达 100%！正在生意人犹豫不决之时，财主又有了一个主意，他想，如果改成半年的利率 50%，还是借一年，那么，半年后可以得到 1.5 两银子，一年后就可以得到 2.25 两银子，这样赚得更多！他赶紧收回了此前的条件，改成了半年还钱的新条件。可是，话刚说完，他就又后悔了。既然半年还钱比一年还钱赚得更多，那为何不改为每月还钱、每周还钱、每天还钱呢？于是财主赶紧回屋拿起笔来算一算。

半年还一次，利率 50%，还钱总数是 $(1+0.5)^2=2.25$（两）；

每月还一次，利率 1/12，还钱总数是 $(1+1/12)^{12}=2.613\ 0$（两）；

每周还一次，利率 1/52，还钱总数是 $(1+1/52)^{52}=2.692\ 6$（两）；

每天还一次，利率 1/365，还钱总数是 $(1+1/365)^{365}=2.714\ 6$（两）。

计算结果让财主十分失望，还钱总数并没有预想的那么多。到这里读者一定看出来了，如果我们把每天再拆成每一小时、每一分钟、每一秒钟，还钱总数会增长的更加缓慢，最终会越来越接近神奇的自然对数底 e。从数学的角度来看，当 x 趋于无穷大时，$(1+1/x)^x$ 的极限值正是 e。

1/e 的值是 0.367 9…，近似为 37%，它与小概率事件之间的神秘关系源于“小概率事件定律”。小概率事件定律，是指一个十分罕见的随机事件，几乎只发生过一次，并且今后能否再次发生难以预测，那么这个事件不再发生的概率是 1/e。被鸟粪砸中、兔子撞树、飞机失事都满足上述条件，因此这些事件不再发生的概率都是 37%。

小概率事件定律听起来有些玄妙，其实背后也是有数学原理的，这就是泊松分布。

泊松分布

被雷劈、中彩票、飞机失事等小概率事件总是让人难以捉摸，它们很少发生，几乎无法预测，即便如此，概率统计还是有办法用数学公式来描述它们。泊松分布正是用来描述那些无法预测的小概率事件发生次数的分布，设随机变量 X 表示某事件发生的次数，若 X 服从泊松分布，则有

$$P(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},\quad k=0,1,2\cdots$$

公式中的 λ（英文写作 lamda）是一个常数，泊松分布的期望和方差都是 λ，图 4-4 是 $\lambda=1$ 时的泊松分布图。

当 $k=0$、$\lambda=1$ 时，$P(X=0)=1/\mathrm{e}$，这便是小概率事件定律的数学原理。

泊松分布在生活和科研中的应用十分广泛。比如每个小时进入银行办理业务的人数、报纸上每一页的错别字数量、某个网页的点击量。网页的点击量？你肯定会对这个例子表示质疑，因为点击某个网页未必是小概率事件，如果这个网页是谷歌、百度的首页怎么办？答案是缩短时间跨度。泊松分布描述的是一个小概率事件在单位时间内发生的次数，这里的“单位时间”是可以任意指定的，对一个热门网页来说，一秒的点击量可能都有上万次，肯定算不上小概率事件，那么我们就把单位时间调整到一毫秒甚至一微秒，在那样的

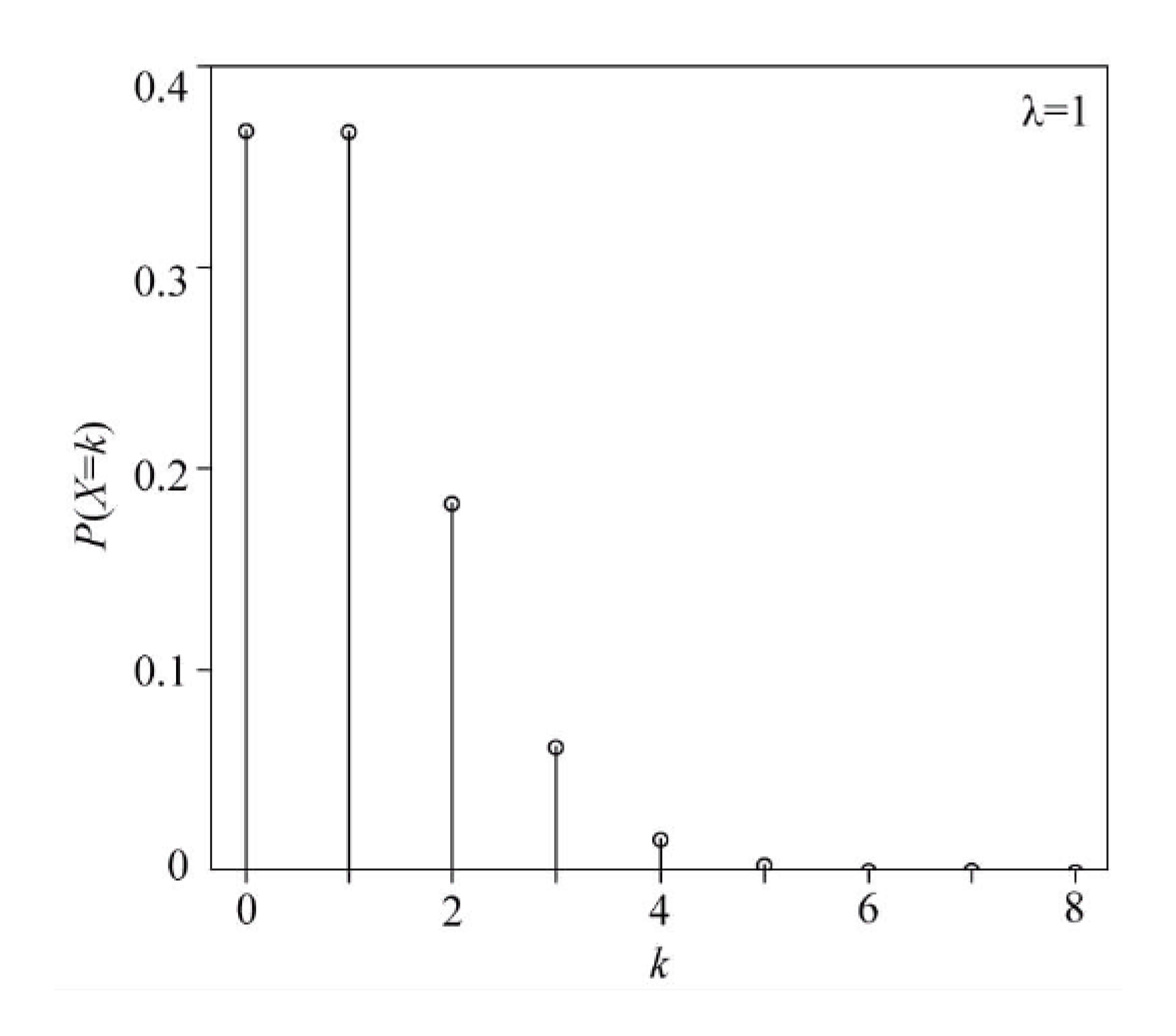


图 4-4 泊松分布

“单位时间”里，网页点击一定可以算作小概率事件了。另外，泊松分布所描述的事件一定是无法预测的随机事件，以网页点击来说，全球几十亿网民随时可能会点击某个网页，如此难以预测的事件一定是随机事件。

回顾泊松分布的表达式，除了自然对数底 e 之外，还有一个常数 λ，这个常数是怎么来的呢？

这需要从二项分布谈起。我们知道，美式大转盘共有 38 个数字，每一局只会出现一个数字，所以每个数字出现的概率都是 1/38。以数字“00”为例，“00”在每一局中出现的概率都是 $p=1/38$，那么，在 $n=38$ 局游戏中，“00”出现 k 次的概率是多少？

我们把每一局的结果分为“00”和“非 00”两种结果，于是，大转盘游戏变成了一个伯努利试验，回顾上一节学习的二项分布，“00”出现 k 次的概率是

$$P(X=k)=C_n^k\cdot(1-p)^{n-k}\cdot p^k=C_{38}^k\cdot\left(1-\frac{1}{38}\right)^{38-k}\cdot\left(\frac{1}{38}\right)^k$$

在这里，我们特意选择了 $n=38$ 局，是因为我们需要 np 成为一个常数，这个常数就是 λ。我们设 $\lambda=np$ 是一个常数，用 λ/n 代替 p，可以得到

$$P(X=k)=C_n^k\cdot(1-p)^{n-k}\cdot p^k$$

$$=\frac{1}{k!}\cdot\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-k-1}{n}\cdot\lambda^k\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

$$\approx\frac{\lambda^k}{k!}e^{-\lambda}$$

泊松分布出现了，它是二项分布的近似表达式。在上面的例子中，$n=38$，$p=1/38$，因此 λ 是 1。我们也可以令 λ 为其他常数，只要你取适合的 n 和 p 就可以了。

在求解概率问题的过程中，如果 $n>20$ 并且 $p<0.05$，我们就可以用泊松分布来近似二项分布，这种近似会帮助我们大大简化计算过程。

4.6　正态分布：完美曲线

电子体重计是当下很多家庭的必备家电，一家人隔三岔五称称体重，各有目标：爸爸要变得健壮，妈妈要保持身材，孩子要茁壮成长。我们用电子体重计时，测一次足矣，虽然初中课本教过我们，“测量有误差，多次测量可以减小误差”，可是我相信，没人会为了消除误差测上五次十次的，除了某些数学天才。

亨利·庞加莱(1854—1912)是法国数学家、天体力学家、数学物理学家、科学哲学家，被公认为 19 世纪末和 20 世纪初的领袖数学家。他有一桩与称重有关的轶事。庞加莱常去住处附近的一家面包店买面包，每次买一块，重量是一千克。不知是出于怀疑还是天生处女座，庞加莱每次买完面包回到家都要再称一次面包重量，然后把重量记在本子上。就这样，庞加莱坚持称重一年，计算出重量的平均值是 950 克，甚至还画出了一个直方图，如图 4-5 所示。然后，他报了警！他举报这个面包店缺斤少两，数据和直方图可以作证。这家倒霉的面包店被迫停业整顿一个月。面包店重新开张后，庞加莱继续买面包，继续称重，继续记录，继续画图。一年以后，庞加莱再次计算面包重量的平均值，结果是 1 千克，看起来面包店改正了自己缺斤少两的问题，可是，庞加莱观察直方图时还是发现了问题：面包的重量本应服从均值为 1 千克的正态分布，可是，直方图的形状明显不符合！庞加莱稍一动脑便猜到了原因：面包店并没有改正缺斤少两的问题，只不过把重一些的面包特意卖给了自己！于是，他又报了警……

在这桩轶事中，最摸不着头脑的非警察莫属，他们要想搞明白庞加莱报警的原因，必须学会概率统计中最重要、最常用的正态分布！

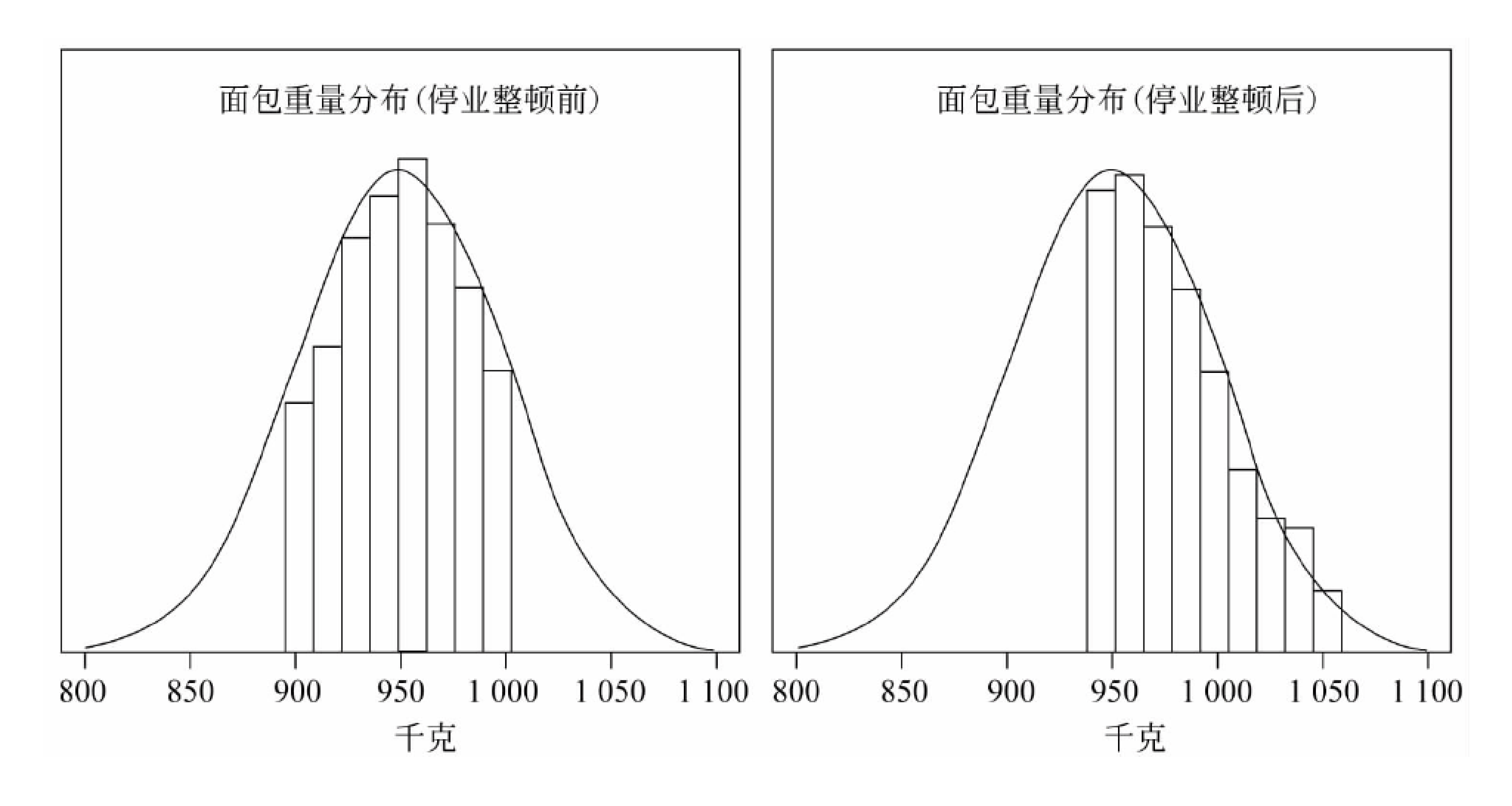


图 4-5 庞加莱绘制的面包重量分布图

正态分布

正态分布，又称高斯分布，是概率统计中最常用的概率分布，与此前学习的概率分布不同，正态分布是连续随机变量的概率分布，在描述连续随机变量的分布时，我们使用概率密度函数 $f(x)$，而不是 $P(X)$，$f(x)$来源于微积分，这里不做详述，读者们可以把 $f(x)$当作 $P(X)$的一种微观表达方式。

如果随机变量 X 的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x-u)^2}{2\sigma^2}}$$

则称 X 服从正态分布。我们不需要记住这个复杂的公式，但一定不能忘记正态分布那条完美的钟形曲线，如图 4-6 所示。

正态分布的期望为 μ，方差为σ^2，标准差为 σ，我们常把期望为 μ、方差为σ^2的正态分布记为 $N(\mu,\sigma^2)$，随机变量 X 服从该分布记为 $X\sim N(\mu,\sigma^2)$。图 4-6 是标准正态分布 $N(0, 1)$的概率分布曲线，从图中可以看出，标准正态分布关于 $x=0$ 左右对称，此外，图 4-6 还标注了随机变量 X 的值落在$[-1,1]$、$[-2,2]$和$[-3,3]$区间的概率大小，X 的值处于$[-3,3]$区间的概率达到了 99.7%，接近 100%！这个特性叫作“3σ 法则”，它可以拓展到所有的正态分布，即服从正态分布 $N(\mu,\sigma^2)$的随机变量的值几乎一定会落在$[\mu-3\sigma, \mu+3\sigma]$这个区间内。

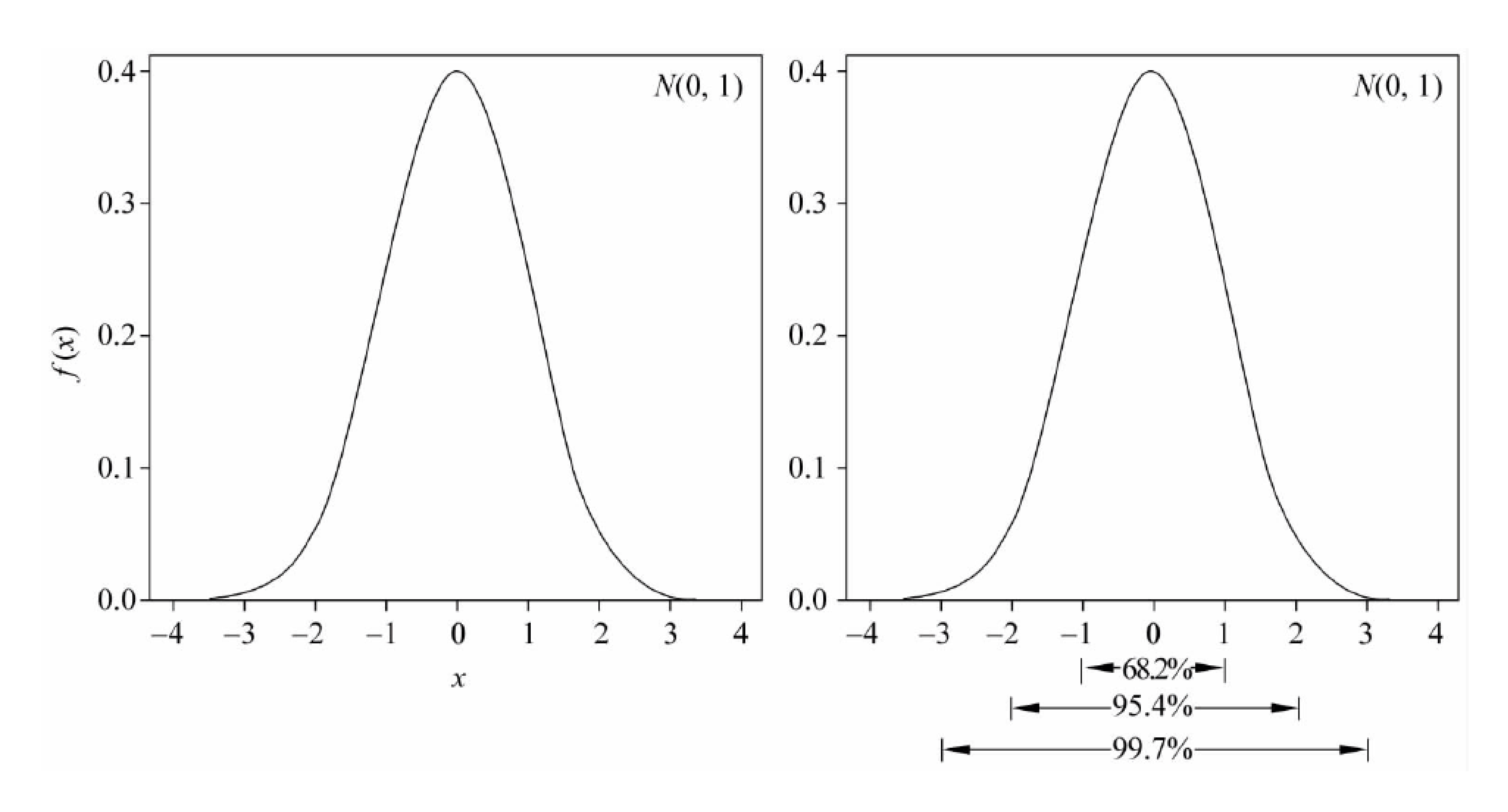


图 4-6　正态分布

在前文中，我们多次提到正态分布是“最常用”的概率分布，这可不是空穴来风，正态分布有一种独一无二的能力——化繁为简。在庞加莱称面包的例子中，庞加莱一口咬定，面包的重量服从正态分布，这是为什么呢？面包虽小，所含的成分却不少，面粉、水分、盐、酵母甚至空气都是面包的成分，每一种成分的重量都有或多或少的随机性，要计算这些随机变量相加之后的概率分布一定十分复杂，大概只有天才数学家才能搞定吧。其实不然，或许庞加莱连面包的成分都不清楚，但他可以确定，面包的重量服从正态分布，因为他懂得——中心极限定理。

中心极限定理是与大数定理并列的重要概率理论，它有几种不同的表达方式，核心思想是，大量的独立随机变量相加，不论各个随机变量的分布是怎样的，它们的加和必定会趋向于正态分布。面包里虽然有很多种未知分布的随机成分，只要这些成分加在一起，一块面包的重量便会服从正态分布。

读者还记得“大数定理”吗？“大数定理”的另一种表达方式是“均值定理”，其含义是，随机变量 X 多个观察值的均值会随着观察值的增加越发趋近于期望值 μ，中心极限定理进一步告诉我们，均值服从期望为 μ 的正态分布。在各种测量试验中，我们一般都认为，测量结果的均值服从正态分布，根据总体均值估计的结论，正态分布的期望 μ 是应与观察值的均值近似相等，这就是庞加莱用来证明面包店缺斤短两的数学原理。

三大分布

正态分布是概率统计最重要的分布，由它演变而来的另外三个分布并称统计学"三大分布"，在统计学中有很广泛的用途，下面我们就来认识一下它们。

X^2 分布

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(0,1)$ 的样本，则称统计量

$$X^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

服从自由度为 n 的 X^2 分布（读作"卡方分布"），记为 $X \sim X^2(n)$，概率分布如图 4-7 所示。

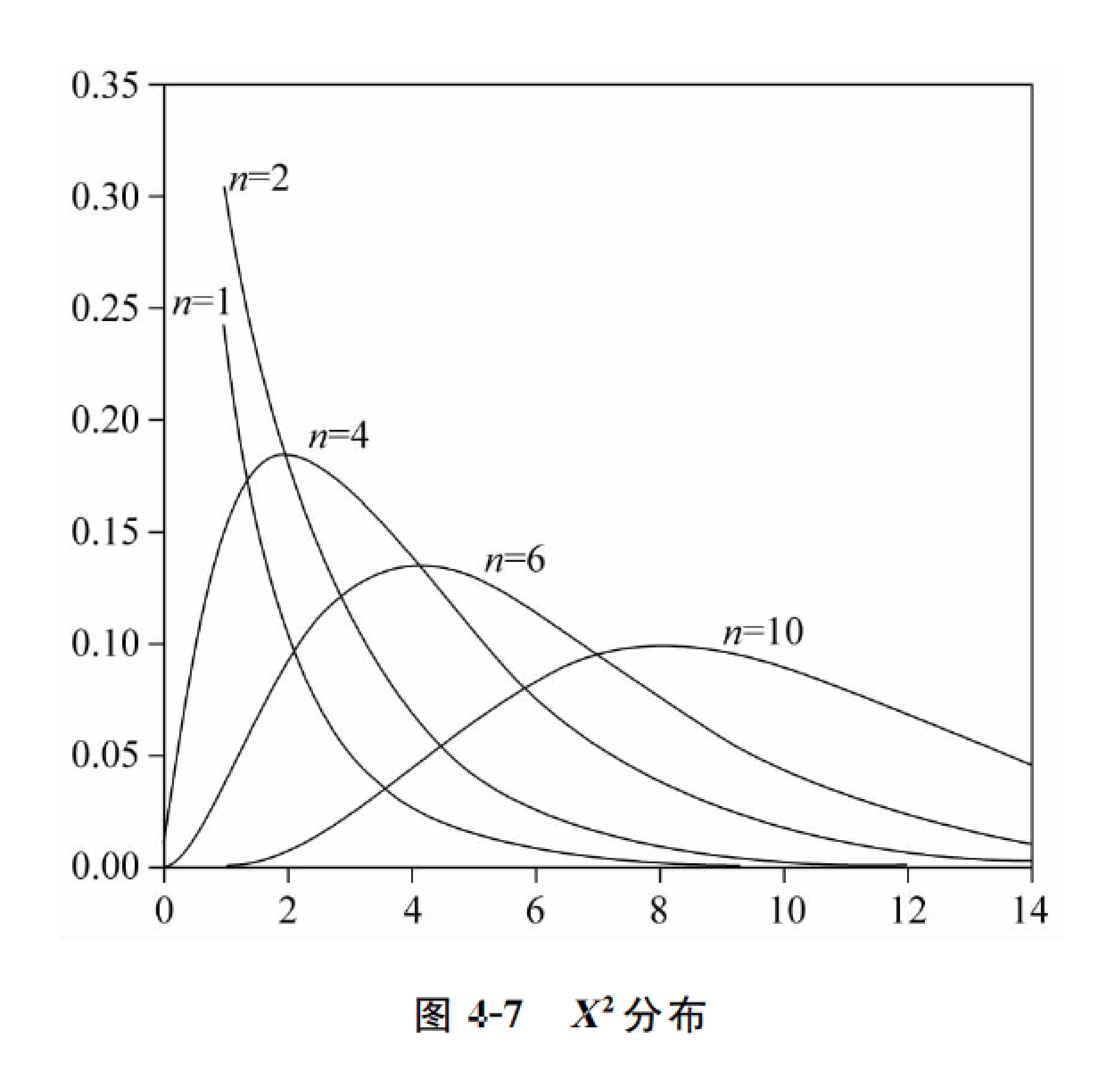


图 4-7　X^2 分布

X^2 分布的期望和方差分别是

$$E(X^2) = n, \quad D(X^2) = 2n$$

t 分布

设 $X \sim N(0,1)$，$Y \sim X^2(n)$，并且 X 和 Y 互相独立，则称随机变量

$$t = \frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布，记为 $t \sim t(n)$，概率分布如图 4-8 所示。

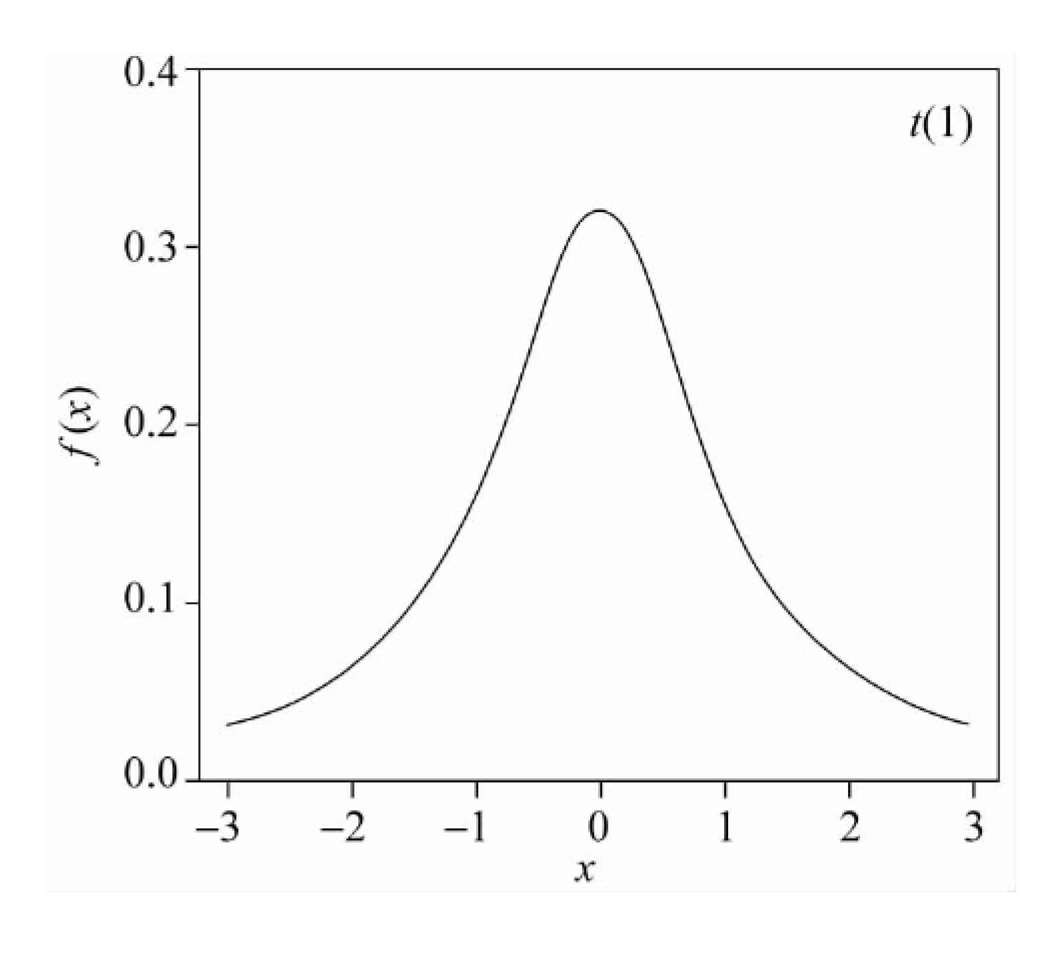


图 4-8　t 分布

F 分布

设 $X \sim X^2(n_1)$，$Y \sim X^2(n_2)$，且 X 和 Y 互相独立，则称随机变量

$$F = (X/n_1)/(Y/n_2)$$

服从自由度为(n_1, n_2)的 F 分布，记为 $F \sim F(n_1, n_2)$，概率分布如图 4-9 所示。

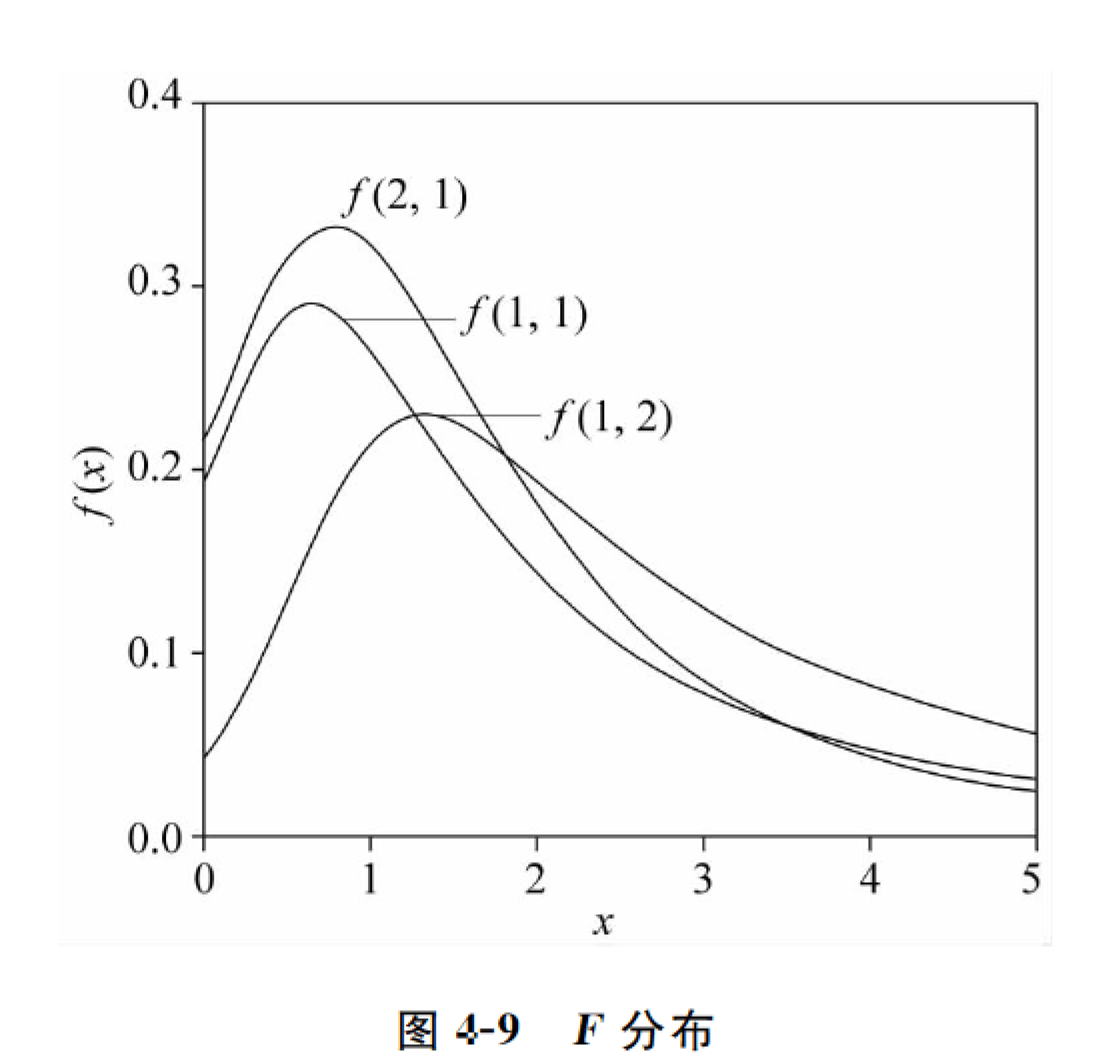


图 4-9　F 分布

这三大分布在假设检验、参数估计等统计学问题中常常使用，本书不会对这三大分布做深入介绍，感兴趣的读者可以阅读统计学的专业书籍。

4.7 指数分布：“二八”与“长尾”

强大的指数

提起指数，读者们一定对“棋盘上放麦粒”的故事很熟悉，这个故事源自古印度的一个古老的传说。舍罕王打算重赏象棋的发明者宰相达伊尔，达伊尔跪在国王面前，提出了自己的请求：“陛下，请您在棋盘上第一个小格里放一粒麦子，第二个小格里放两粒麦子，第三个小格里放四粒麦子，如此这般，直到填满整个棋盘，这就是微臣要的奖赏。”国王一听，觉得这样的要求实在不足为奇，但既然达伊尔如此要求，便下令满足他的要求。仆人们扛来一袋麦子，本以为足够，可是还没填满十格就不够了，之后，一袋又一袋的麦子被扛过来，距离填满棋盘依然遥遥无期。最后，国王不得不承认，倾全国之麦粒也无法满足达伊尔的请求。

国际象棋的棋盘有 64 个格，按照达伊尔的请求，最后一个格子里要放 2^{63} 粒麦子，我们用计算机的常用计量单位来衡量这个数字，2^{13} 大约是 1KB，2^{23} 是 1MB，2^{33} 是 1GB，2^{43} 是 1TB，2^{53} 是 1PB，2^{63} 是 1EB，即使对超级计算机来说，这也是个十足的“大数据”！

指数既能把数字变得无穷大，也能把数字变得无穷小。有这样一个与指数有关的问题：假设有一种细胞，分裂和死亡的概率相同，都是 50%。如果一个物种从这样一个细胞开始进化，那么这个物种灭绝的概率是多少？

直觉告诉我们，应该是 50%吧。细想想，如果细胞一开始就死亡，物种便灭绝了，概率是 50%；如果第一个细胞分裂为两个细胞，这两个细胞有可能全部死亡，这种情况的概率是 50%×50%×50%=12.5%，如此一直计算下去，会得到无穷多的概率，这些概率相加就是物种灭绝的概率。因此这个概率肯定大于 50%，可是究竟是多少，我们来算一算。

设 A 表示物种灭绝事件，B_1 表示第一个细胞分裂，B_2 表示第一个细胞死亡，根据全概率公式，有如下等式：

$$P(A) = P(A \mid B_1) \cdot P(B_1) + P(A \mid B_2) \cdot P(B_2)$$

很显然，$P(B_1)=50\%$，$P(B_2)=50\%$，$P(A|B_2)=1$。$P(A|B_1)$表示第一个细胞分裂的前提下，物种灭绝的概率，第一个细胞会分裂为两个独立的细胞，因此 $A|B_1$事件等同于“两个细胞各自分裂或死亡，最终物种灭绝的概率”，由于这两个细胞彼此独立，因此“两个细胞导致物种灭绝”的概率是“一个细胞导致物种灭绝”的概率的平方，即 $P(A|B_1)=[P(A)]^2$，这与编程中的递归算法异曲同工。

我们用 p 代替 $P(A)$，便可以得到如下等式：

$$p = p^2/2+1/2$$

解这个方程，会得到一个惊人的答案：$p=1$，物种必然会灭亡！

这就是指数，不论变大还是变小，它总是拥有无比强大的爆发力！

指数分布

在概率统计中，也存在一个与指数有关的分布——指数分布。

如果随机变量 X 的概率密度函数为

$$f(x)=\begin{cases}\dfrac{1}{a}\mathrm{e}^{-x/a}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

则称 X 服从参数为 a 的指数分布，其中 a 为大于 0 的常数。

图 4-10 为 a 取不同数值时的指数分布曲线。

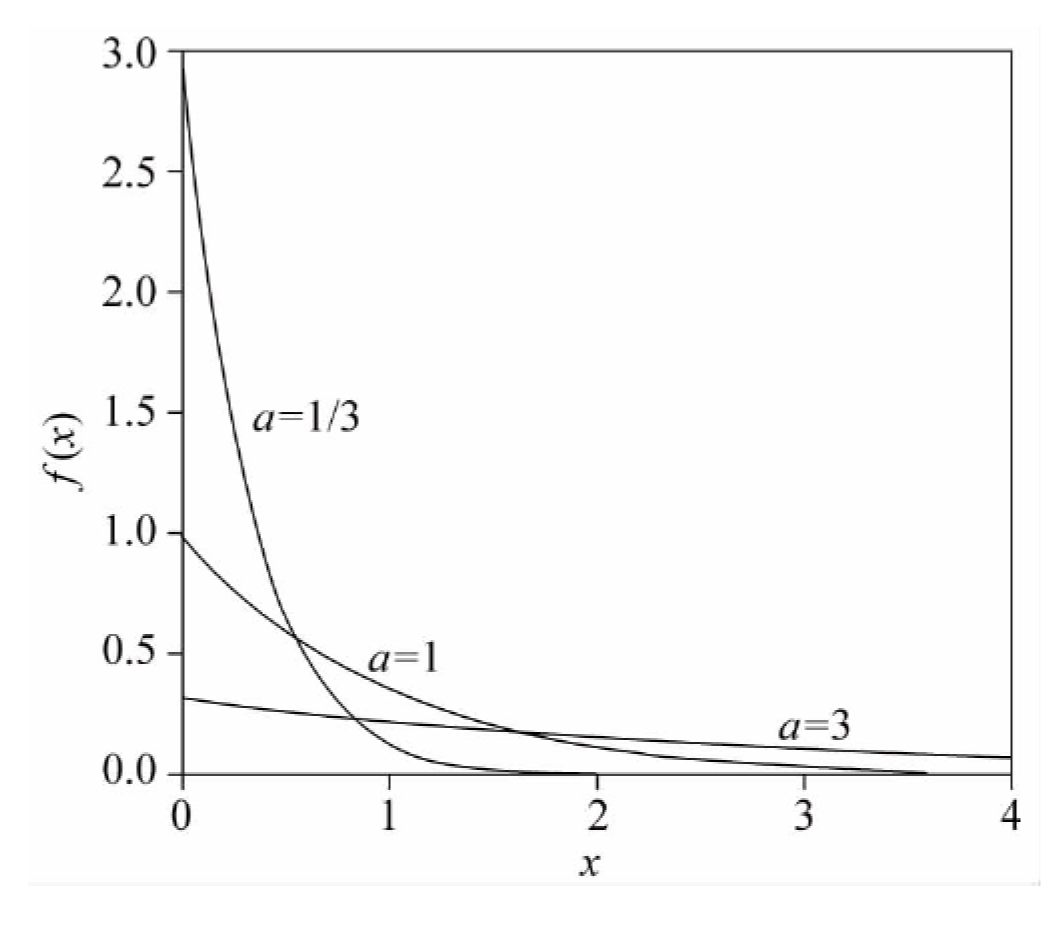


图 4-10　指数分布

指数分布的一个重要的性质是“无记忆性”，它指的是服从指数分布的随机变量 X 满足：

$$P(X > s+t \mid X > s) = P(X > t)$$

其中，s 和 t 是两个常数。

举个例子，设随机变量 X 是灯泡的使用时间，X 服从指数分布。那么，上面的等式可以解读为，灯泡在已经使用 s 小时的条件下，使用时间长于 $s+t$ 小时的概率与灯泡使用时间长于 t 小时的概率是相等的，看起来，灯泡似乎“忘记”了自己曾经使用了 s 小时，这就是“无记忆性”，正因为这一特性，指数分布常常应用于排队论中。

生活在人来人往的社会中，排队是每天必做的事情。上下班排队等公交车、去超市购物排队交费、开车出游排队过收费站、牵着爱人的手到民政局也要排队领结婚证。排队论，也称随机服务系统理论，它通过对服务对象到来及服务时间的统计研究，得出统计规律，再根据这些规律来改进服务系统的结构。

我们以银行为例来说明排队论原理。银行一般会开设若干窗口为顾客服务，顾客依次进入大厅，刷卡领号，然后坐在大厅中等候叫号，这是一个非常典型的排队论研究场景。排队论中常常假定顾客的到来是“不可预测”的随机事件，所以顾客单位时间内到达的人数服从泊松分布，与之相对应的，顾客的到达时间间隔恰恰服从指数分布，我们设单位时间内到达的顾客数量为 λ，则顾客的到达时间间隔 T 服从如下的概率密度函数：

$$f(t) = \lambda e^{-\lambda t}, \quad t \geqslant 0$$

式中，T 的均值为 $1/\lambda$，方差为 $1/\lambda^2$。

指数分布的无记忆性体现在，从任意时刻算起，顾客的到达时间间隔都服从同样的指数分布，这正是指数分布的神奇之处。另一个典型的排队论场景是排队等待公交车。在交通繁忙的城市里，公交车的到站时间往往难以预测，因此公交车的到达时间间隔也近似服从指数分布，这就意味着，无论你什么时候到达车站，等候时间都服从同样的指数分布。所以，刚刚错过一辆未必意味着需要等待很久，已经等了很久未必意味着车会马上来，在公交车站里，我们能做的只有耐心等待。

“二八定律”与“长尾理论”

由指数分布衍生出了两个著名的理论——“二八定律”和“长尾理论”。

“二八定律”指的是生活中的许多不平衡现象往往呈现 20％、80％的分布规律，比如，社会上 80％的财富被 20％的富人占有，公司 80％的收益来自 20％的客户，行业里 80％的市场份额被 20％的强势品牌所占有。

“二八定律”又称帕累托定律，它源自意大利经济学者帕累托的一个发现。1897 年，帕累托偶然注意到 19 世纪英国人的财富和收益模式。在调查取样中，他发现大部分财富流向了少数人手里，这种微妙关系在其他国家一再出现，而且在数学上呈现出一种稳定的关系。帕累托从中总结出这样的规律：财富在人口中的分配是不平衡的，社会上 20％的人占有 80％的财富。

在帕累托定律之后，人们相继发现很多领域都存在类似的不平衡现象。一个知名的例子是犹太人经商的“二八定律”。美国企业家威廉·穆尔在为格利登公司销售油漆时，第一个月仅挣了 160 美元。此后，他学习犹太人经商的“二八定律”，分析自己的销售图表，发现 80％的收益来自 20％的客户，但是他却对所有客户花费了同样多的时间。于是，威廉·穆尔把最不活跃的 36 个客户分派给其他销售人员，自己则把精力集中到那 20％的客户上，不久，他一个月就赚到了 1 000 美元。威廉·穆尔从此学会了犹太人经商的“二八定律”，连续九年坚持这一法则，最终成为凯利—穆尔油漆公司的董事长。

“二八定律”是对线性思维的颠覆，它提醒我们，财富的分配往往是不平均的，因此，我们也不应该用简单的线性思维来分配我们的时间，把更多的时间用在最有成效的“20％”身上才能走上成功之路！

互联网时代催生的“长尾理论”是对“二八定律”的颠覆，与“二八定律”相反，“长尾理论”关注指数分布的后 80％的“利基市场”，如图 4-11 所示。“长尾理论”认为，在高度互联的网络时代，商品的生产、存储、流通、销售的成本大大降低，需求量较低的小众产品可以毫不费力地找到买家，大量的小众产品会累积成很大的市场份额，甚至可能超过那 20％的主流产品。例如，一家大型书店通常可摆放 10 万本书，但亚马逊网络书店的图书销售额中，有 1/4 来自排名 10 万以后的书籍，而且这一比例仍在上升。

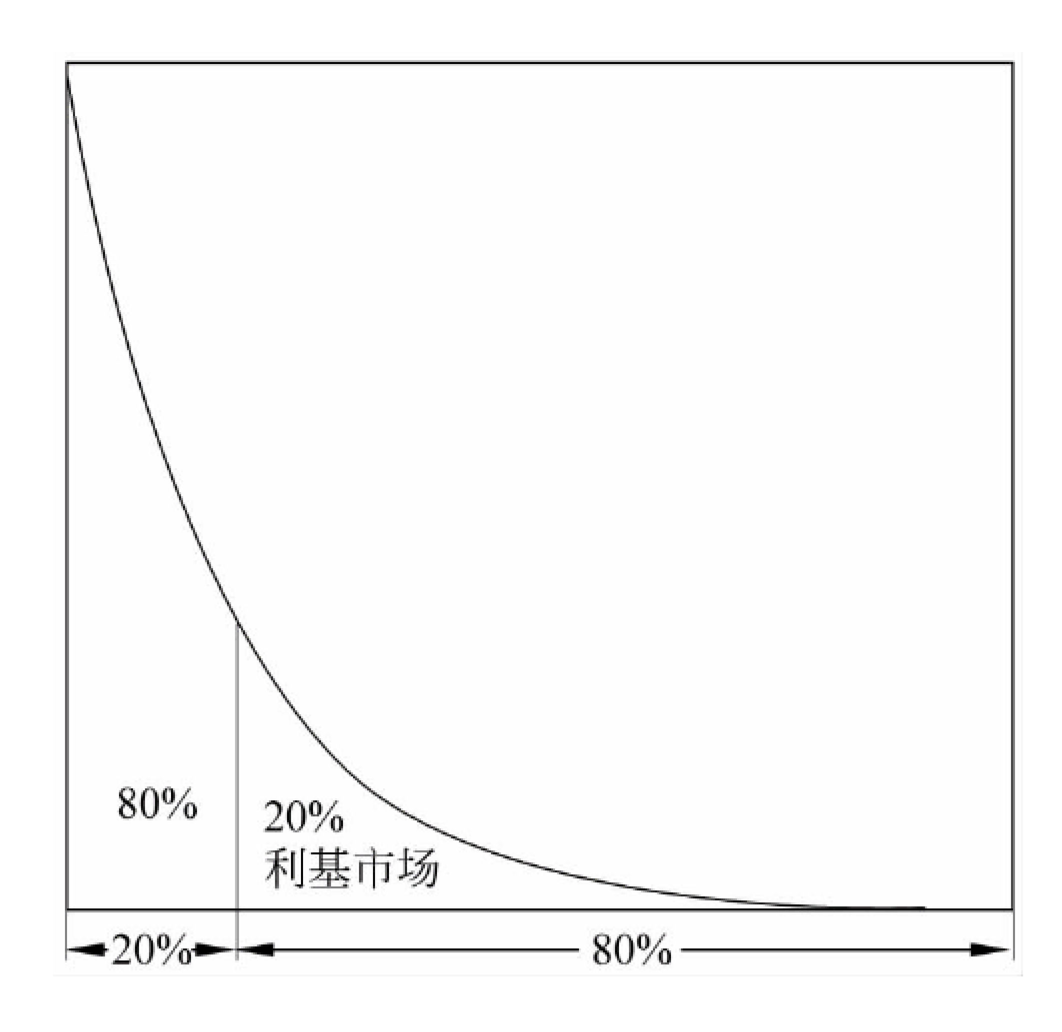


图 4-11 “二八定律”示意图

“长尾”一词最早由美国《连线》杂志主编克里斯·安德森提出。克里斯·安德森喜欢从数字中发现趋势，在跟 eCast 首席执行官范·阿迪布的一次会面时，阿迪布提出一个让人耳目一新的“98 法则”——数字音乐的点唱统计结果显示，听众对 98%的非热门音乐有着无限的需求，非热门音乐的潜在市场空间无比巨大。安德森意识到，这个有悖常识的“98 法则”或许隐含着一个真理。于是，他系统研究了 Amazon、Google、eBay、Netflix 等互联网零售巨头的销售数据，并与沃尔玛等传统零售商的销售数据进行了对比，得到了一条需求曲线，这条曲线拖着长长的尾巴，“长尾”由此得名。安德森把他的发现整理成文章，标题正是“长尾”，这篇文章刊登在《连线》杂志 2004 年 10 月号，后迅速蹿升为这家杂志历史上被引用最多的文章，随后安德森据此撰写了《长尾理论》，这本书也一举登上纽约时报畅销书排行榜。



第 5 章

赌博中的概率统计

导语：赌博，永远不缺乏激情，可是鲜为人知的是，赌博的原理恰恰是严谨的概率统计。学会了赌博中的概率统计，可以让我们更加享受赌局，真正做到“充满激情的同时不丧失理性，充满理性的同时不丧失激情”！

5.1 赌博：激情与理性

在我们的意识里，赌博总被认为是低级、负面的，然而细细想来，赌博本是一个中性词。中国澳门赌场里的老虎机、大转盘是赌博，过年时家里摆上两桌麻将也是赌博，对大多数人来说，赌博是为了体验“未知”带来的刺激，就像球迷们盯着电视看点球大战一样，那千钧一发的时刻总是充满变数，无比刺激！

赌博早在几百年前就已经存在，如今甚至发展成为一个独立的产业——博彩业，世界四大赌城——拉斯维加斯、大西洋城、蒙特卡洛和中国澳门——正是博彩业的象征。

博彩业有很多分支，比如彩票、赌场、赛马等，我国的博彩业主要由福利彩票和体育彩票构成。

中国福利彩票始于1987年，以“扶老助残，济困救孤”为宗旨，包括了刮刮乐、双色球、35选7等多种数字型彩票。中国体育彩票是竞猜体育比赛结果的彩票，涵盖足球、篮球等多个体育项目，其中足球彩票的发行量最大，玩法最多。

赌场中的赌博花样繁多，老虎机、大转盘是赌博机的代表，得州扑克、21点是扑克类的代表。

赛马，又称赌马，是对跑马结果进行竞猜的一种彩票。由巴黎实业家奥莱于19世纪末发明，后来成为全世界最盛行的一种赌博，现在在中国香港非常流行。

无论哪一种博彩方式，都建立在概率统计的基础之上。在前面第2章中我们提到，概率论起源于骰子游戏的研究。伽利略、帕斯卡、费马等多位数学家都曾研究过骰子游戏中的概率问题，“样本空间”“条件概率”等概念也从这些研究中萌芽出来。后来，概率论形成并逐渐完善，催生了博彩种类的丰富，铸就了博彩业的兴盛。一个合格的博彩玩家必须懂得博彩背后的概率原理，否则，他一定是赌局里那个头脑发昏的笨蛋！

总而言之，要真正享受博彩的乐趣，就要做到“充满激情的同时不丧失理性，充满理性的同时不丧失激情”！本章，我们就来聊聊隐藏在赌局背后的概率统计原理。

博彩业各种玩法示意图如图5-1所示。



图5-1　博彩业各种玩法示意图

5.2　双色球：千年等一回

数字类彩票规则简单、操作方便，是全球最流行的博彩方式。我国的数字型彩票种类繁多，包括双色球、排列五、排列三、刮刮乐、35 选 7 和各种地方福利彩票。下面，我们以双色球为例，一起来学习数字型彩票的概率原理。

投注规则

双色球是我国数字型彩票的经典彩种，于 2003 年开始在全国联网发售，是现在全国销售额最大的彩种之一，曾经出现过多位奖金过亿的中奖者。

双色球的投注规则是：双色球投注区分为红球号码区和蓝球号码区，红球号码区由 1～33 共 33 个号码组成，蓝球号码区由 1～16 共 16 个号码组成。投注时选择 6 个红球号码和 1 个蓝球号码组成一注进行单式投注，如图 5-2 所示，每注金额人民币 2 元。



图 5-2　双色球示意图

单式投注：规则中的“单式投注”是指投注者每次只选择一组投注号码，例如，红球的 01、02、03、04、05、06 和蓝球 07，或者红球 15、08、13、14、03、30 和蓝球 04（如图 5-2 所示）。

复式投注：与单式投注相对的是复式投注。复式投注是指，投注者一次选择多个投注号码，一次性购买这些号码构成的所有可能的投注，例如，投注者复式投注红球 01、02、03、04、05、06、07，这意味着，投注者将一次性购买由 01、02、03、04、05、06、07 中任意 6 个构成的所有投注号码，包括“01、02、03、04、05、06”“01、02、03、04、05、07”“01、02、03、04、06、07”等。

倍投：成倍投注的简称，指的是投注者对同样的投注号码进行重复购买，

例如,对红球的01、02、03、04、05、06和蓝球07这组号码进行5倍投,意味着投注者购买了5组同样的号码。

双色球共设六个中奖等级,规则如下:

一等奖:投注号码与当期开奖号码全部相同,奖金浮动;

二等奖:投注号码与当期开奖号码中的6个红色球号码相同,奖金浮动;

三等奖:投注号码与当期开奖号码中的任意5个红色球号码和1个蓝色球号码相同,奖金3 000元;

四等奖:投注号码与当期开奖号码中的任意5个红色球号码相同,或与任意4个红色球号码和1个蓝色球号码相同,奖金200元;

五等奖:投注号码与当期开奖号码中的任意4个红色球号码相同,或与任意3个红色球号码和1个蓝色球号码相同,奖金10元;

六等奖:投注号码与当期开奖号码中的1个蓝色球号码相同,奖金5元。

其中,一等奖和二等奖的奖金与每期的彩票销售总额和中奖人数有关,属于浮动型奖金,我们常说的“大奖500万”只是一等奖奖金的一个代称。

投注策略

下面,我们来计算双色球的中奖概率。双色球是一个典型的组合问题,红球是从33个数字中选出6个,蓝球是从16个数字中选出1个,并且红球和蓝球之间互相独立。我们假设投注者购买了一组投注号码,那么,在开奖之前,这组号码的中奖概率分别是:

$P(\text{中一等奖}) = 1/(C_{33}^{6} \cdot C_{16}^{1}) = 0.000\,005\,6\%$

$P(\text{中二等奖}) = C_{16}^{1}/(C_{33}^{6} \cdot C_{16}^{1}) = 0.000\,090\%$

$P(\text{中三等奖}) = (C_{6}^{5} \cdot C_{27}^{1})/(C_{33}^{6} \cdot C_{16}^{1}) = 0.000\,91\%$

$P(\text{中四等奖}) = (C_{6}^{5} \cdot C_{27}^{1} \cdot C_{16}^{1} + C_{6}^{4} \cdot C_{27}^{2})/(C_{33}^{6} \cdot C_{16}^{1}) = 0.044\%$

$P(\text{中五等奖}) = (C_{6}^{4} \cdot C_{27}^{2} \cdot C_{16}^{1} + C_{6}^{3} \cdot C_{27}^{3})/(C_{33}^{6} \cdot C_{16}^{1}) = 0.81\%$

$P(\text{中六等奖}) = 1/C_{16}^{1} = 6.25\%$

这组号码不中奖的概率是:

$$P(\text{未中奖}) = 1 - P(\text{中一等奖}) - P(\text{中二等奖}) - P(\text{中三等奖}) - P(\text{中四等奖}) - P(\text{中五等奖}) - P(\text{中六等奖})$$

$$=92.90\%$$

双色球奖项、中奖概率和奖金如表 5-1 所示。

表 5-1　双色球奖项、中奖概率和奖金

奖项	中奖概率(%)	奖金(元)
一等奖	0.000 005 6	浮动
二等奖	0.000 090	浮动
三等奖	0.000 91	3 000
四等奖	0.044	200
五等奖	0.81	10
六等奖	6.25	5
未中奖	92.90	0

双色球一共有约 1 770 万($C_{33}^{6} \cdot C_{16}^{1}$)组可能的号码,要中一等奖,需要所有号码都相同,因此中奖的概率便是 1 770 万分之一,即 0.000 005 6%。双色球每周销售三期,一年有 52 周,因此,一年里我们可以投注双色球约 150 次,如果每次单式投注一组号码,中一次一等奖平均需要 11.8 万年,即使每次投注 100 组号码,平均也需要 1 180 年,真可谓“千年等一回”,想要战胜小概率事件谈何容易!

假定一等奖的奖金为 500 万元,二等奖的奖金为 50 万元,一组单式投注号码的奖金期望是:

$$\begin{aligned} E(\text{奖金}) =& 5\ 000\ 000 \times P(\text{中一等奖}) + 500\ 000 \times P(\text{中二等奖}) + 3\ 000 \times \\ & P(\text{中三等奖}) + 200 \times P(\text{中四等奖}) + 10 \times P(\text{中五等奖}) + \\ & 5 \times P(\text{中六等奖}) \\ =& 1.24(\text{元}) \end{aligned}$$

每组号码的投注金额是 2 元,因此,一组号码的收益期望为:

$$E(\text{收益}) = 2 - 1.24 = -0.76(\text{元})$$

投注者每购买一组号码,平均会损失 0.76 元。

现在,我们知道了两件事:一是中一等奖的概率非常低;二是买双色球不可能赚到钱。其实,这两件事是众所周知的,我们只是用数学算式验证了它们是正确的。那么,投注者为什么还要买彩票呢?为了那看似渺茫的中奖机会!不管中奖概率有多低,总有人中大奖,所以我们还是要买,而且要坚持买!那么,怎么买才更合理呢?或者换一种问法,有没有什么方法能提高中奖概率?

据我的观察，买双色球的人大约用三种方法选择号码：第一种是机选，你只要走到投注网点，掏出 2 元钱，来一注机选，投注设备会随机帮你选出一组号码；第二种是守号，你躺在床上冥思苦想出一组号码，里面可能包含你的生日和你的幸运数字，你觉得这组号码是属于你的独家搭配，于是你每次都买这组号码，高兴了还会来个倍投；第三种是自主选号，每当买彩之前，你都苦思冥想一阵子，神秘的第六感指引你写出一组号码，就买它了！

这三种方法的区别在于两点：一是号码由机器选出还是你自己选出；二是每次买的号码相同还是不同。这三种方法的共同点只有一个，你会在开奖之前买一组号码，而且不可更改。不论我们在彩票打印出来之前做了什么，我们都会花 2 元钱，买一组号码，当这组号码已经确定之时，一切的选号方法都没有意义了，你只能坐等晚上九点半的开奖。不论你是怎么选出这组号码的，也不论这组号码是什么，此时此刻，你中大奖的概率就是 0.000 005 6%，不会更大，也不会更小。

总有些人不死心，因为他们相信“大数定理”，就连彩票网站上也会提供类似如图 5-3 所示的号码走势图。在“大数定理”一节，我们已经讨论过，大数定理并不会使硬币的正反两面出现的次数越来越接近，即使连续十次都是正面，我们依然认为第十一次出现正面的概率是 50%，因为每一次抛掷是独立的。同理，双色球中33 个红球和 16 个蓝球被选出的概率也是相同的，不同期的选号过程也是互相独立进行的，所以，研究号码走势纯粹是在浪费时间！浪费时间！浪费时间！

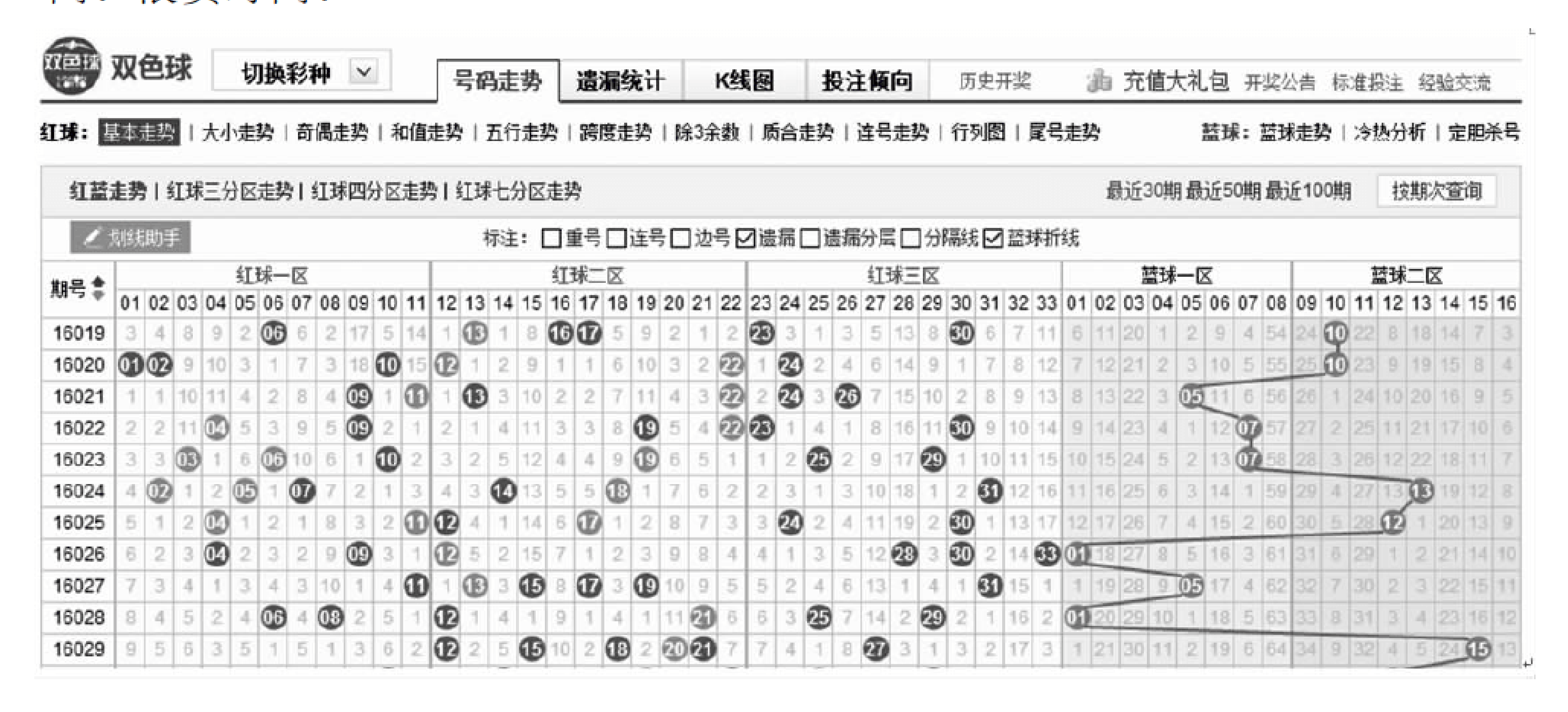


图 5-3　双色球号码走势图

"双色球"小结：

（1）一等奖中奖概率极低；

（2）坚持买双色球不可能赚到钱；

（3）选号方法对中奖概率没有任何影响；

（4）研究号码走势没有意义；

（5）切记小赌怡情。

5.3　足彩：爱足球，更爱足彩

1998 年法国世界杯让我爱上了足球，从圣西罗到威斯特法轮，从西班牙国家德比到英超双红会，我一直是欧洲联赛的忠实观众。看球久了，自然喜欢上了猜球——猜胜负、猜比分，猜球终归不过瘾，就开始买足彩。

足球彩票，简称足彩，是起源于欧洲的体育类彩种。在欧美地区，足彩由合法注册的博彩公司负责销售。我国的足彩起步较晚，于 2001 年 10 月上市，由中国体彩中心负责销售。从 2001 年至今，足彩的玩法几经变化，现行的玩法包括 14 场胜负彩、任选 9 场胜负彩、进球彩、半全场等。与数字型彩票相比，足彩包含的元素要丰富得多，赛前有关球队的打法、状态、心态、赔率甚至花边新闻都是玩家们关心的话题，而且一场足球比赛有 90 分钟，在如此长的"开奖时间"里，比赛结果随时可能会变化，补时阶段的一个进球，既可能让你喜中大奖，也可能使你与大奖失之交臂，这就是足彩令人着迷之处。

我曾短暂地沉迷于足彩，还中过两次小奖，但是我对足彩的认识一直停留在感性层面，在学习了概率统计后，我尝试着用概率统计的方法分析足彩，接下来，我以 14 场胜负彩为例，与大家分享我的足彩心得。

投注规则

14 场胜负彩的投注规则是：以 14 场比赛的最终结果进行投注，每场比赛的结果分为"胜、平、负"三种，"胜"表示主场球队取胜，"平"表示两队打平，"负"表示主场球队告负。例如，第 16069 期胜负彩共竞猜 14 场比赛，如图 5-4

所示，每场比赛挑选一个结果，构成一组投注，每注金额人民币2元。

编号	赛事	比赛时间	主队 VS 客队	□胜	□平	□负
1	英 超	05-01 21:05	曼 联 VS 莱切斯特	3	1	0
2	英 超	05-01 23:30	南安普敦 VS 曼 城	3	1	0
3	意 甲	05-01 21:00	AC米兰 VS 弗洛西诺	3	1	0
4	意 甲	05-01 21:00	恩波利 VS 博洛尼亚	3	1	0
5	意 甲	05-01 21:00	巴勒莫 VS 桑 普	3	1	0
6	意 甲	05-01 21:00	萨索洛 VS 维罗纳	3	1	0
7	意 甲	05-02 02:45	拉齐奥 VS 国际米兰	3	1	0
8	西 甲	05-01 22:00	西班牙人 VS 塞维利亚	3	1	0
9	西 甲	05-02 00:15	拉 科 VS 赫塔费	3	1	0
10	西 甲	05-02 02:30	巴伦西亚 VS 比利亚雷	3	1	0
11	法 甲	05-01 23:00	昂 热 VS 马 赛	3	1	0
12	瑞典超	05-01 21:00	法尔肯堡 VS 埃夫斯堡	3	1	0
13	瑞典超	05-01 23:30	哈马比 VS 松兹瓦尔	3	1	0
14	瑞典超	05-01 23:30	马尔默 VS 赫 根	3	1	0

图5-4　第16069期胜负彩场次

14场胜负彩设置两个奖项：

一等奖：猜中全部14场比赛的胜平负结果，浮动奖金；

二等奖：猜中其中13场比赛的胜平负结果，浮动奖金。

下面，我们来计算14场胜负彩的中奖概率。

每场比赛有“胜、平、负”三种结果，因此，假设猜中的概率为1/3。14场胜负彩是一个典型的14重伯努利试验，每一场比赛就是一次试验，因此，我们可以应用二项分布来计算中奖概率。

一等奖要求14场全部猜中，所以中奖概率为：

$$P(\text{中一等奖}) = (1/3)^{14} = 0.000\,021\%$$

二等奖要求猜中13场，所以，中奖概率为：

$$P(\text{中二等奖}) = C_{14}^{13} \times (1/3)^{13} \times (2/3) = 0.000\,59\%$$

计算结果说明，14场胜负彩的中奖概率非常低，与数字型彩票相似。

投注技巧

虽然足彩的中奖概率很低，但相比于完全随机的数字彩，足彩是可以利用

一些技巧来提高中奖概率的。

技巧一：学会看赔率

足球赔率分为欧洲赔率和亚洲赔率两种。

欧洲赔率的一般形式是：

皇马 VS 拜仁 2.25　3.00　3.25

这是一场欧洲冠军杯比赛——皇马主场对阵拜仁的欧洲赔率，其中的三个数字 2.25、3.00 和 3.25 依次表示胜、平、负的赔率，这三个赔率的含义是：

假如你投注 100 元赌皇马胜，皇马果真取胜，你会得到 225 元（含本金），否则你输掉 100 元；

假如你投注 100 元赌两队打平，两队果真打平，你得到 300 元（含本金），否则你输掉 100 元；

假如你投注 100 元赌皇马输球，皇马果然输球，你得到 325 元（含本金），否则你输掉 100 元。

一般来说，胜、平、负三个结果中，赔率最低的是博彩公司最看好的结果。

亚洲赔率的一般形式是：

皇马 VS 拜仁　让　平手/半球

仍然是皇马对阵拜仁的比赛，亚洲赔率给出的赔率是“主队让平手或半个球”。在亚洲赔率中，看好主队则“让”，看好客队则“受让”，在“让”或“受让”后边，会出现“平手/半球”“半球”“一球”“一球半”等，表示赔率的大小。例如，“让一球”表示主队至少赢客队一个球，“让两球”表示主队至少赢客队两个球，“让两球”比“让一球”更能展现出赔率对主队获胜的信心。除了让球，亚洲赔率中还有贴水，用于计算奖金，类似于欧洲赔率中的三个数字，此处不再详述。此外，投注者一定要注意，各个博彩公司的赔率会随时变化，直到比赛结束，投注者只需把赔率当作两队实力对比的参考指标，不必刻意关注其中的细微变化。

技巧二：学会实力分析

如果我问你：今晚西甲联赛，皇家马德里队主场对阵希洪竞技队，你认为谁会赢？不管懂球还是不懂球，你一定都想，皇马这么强，怎么会不赢？但是，足球场上，一切皆有可能！要真正提高猜中的概率，就要学会对两支球队做实

力分析，从综合实力、竞技状态、求胜欲望、历史交锋战绩、关键球员伤停等诸多方面来分析两支球队，然后才能做出更加准确的判断。比如，2015－2016 赛季的英超联赛中，上赛季冠军切尔西队表现糟糕，仅仅排在联赛中游，赛季初更是一度掉入降级区，而阿森纳队一如既往地处在联赛前四名。但当两队相遇时，笑到最后的依然是切尔西队，仿佛是两支球队近年来多次交锋的重演，这就是球风相克的典型代表。实力分析包含很多方面，对不同的比赛，我们要分清优先级，有时历史战绩更重要，有时球队竞技状态更关键，其中技巧留作足彩投注者们仔细品味吧。

技巧三：正确理解“冷门”

所谓冷门，就是出人意料的比赛结果，比如，2015－2016 赛季的西甲联赛中，巴塞罗那队在主场 1∶2 负于排名中游的瓦伦西亚队，这就是个大冷门。回顾足彩的历史记录，我们不难发现，冷门似乎常常会发生，这看似不正常的现象其实有合理的解释。经统计，强弱差距很大的比赛，强队取胜的概率可达 70％，每一期足彩的 14 场比赛中，往往有 3～5 场这样的比赛，以 3 场为例，不出现冷门需要三支强队同时取胜，其概率为：

$$P(\text{三支强队同时获胜}) = (70\%)^3 = 34.3\%$$

出现一场冷门的概率为：

$$P(\text{一支强队未取胜}) = C_3^1 \times (70\%)^2 \times (30\%) = 44.1\%$$

出现两场冷门的概率为：

$$P(\text{两支强队未取胜}) = C_3^2 \times (70\%) \times (30\%)^2 = 18.9\%$$

三场均出现冷门的概率为：

$$P(\text{三支强队均未取胜}) = (0.3)^3 = 2.7\%$$

对比上面的结果可以发现，出现冷门的概率（65.7％）比不出现冷门的概率（34.3％）要高得多，而且，出现一场冷门的概率最高。所以，14 场比赛中常会出现一场甚至两场冷门，这正是二项分布的神奇之处！

将上述结论推而广之，可以得到两个推论：一是如果强弱分明的比赛有 4 场、5 场甚至更多，冷门不出现的概率会更低；二是在强队获胜概率为 70％的假设下，不论强弱分明的比赛有 3 场、4 场还是 5 场，出现一场冷门的概率都是最高的。读者可以验证一下这两个推论是否正确。

既然冷门很可能会发生，刻意选择冷门结果更合理吗？要回答这个问

题，我们首先要知道，足彩的奖金是由中奖的投注平均分配的，同样是 1 000 万元的奖金总额，如果有 5 注彩票中奖，则每注奖金 200 万元，如果有 50 注中奖，则每注奖金20 万元。下面，我们就来算一算，什么情况下选择冷门更合理。

假设彩民正在进行单场比赛竞猜，比赛结果分为胜、平、负三种，投注共计 100 注，由于主队实力远胜于客队，其中 90 注选择胜，5 注选择平，5 注选择负，总奖金为 100 元，由猜中者平均分配。在下列两种假设条件下，计算三种投注的奖金期望值：

条件 1：强队取胜、打平和告负的概率为 90％、5％和 5％；

条件 2：强队取胜、打平和告负的概率为 70％、15％和 15％。

根据奖金分配规则，胜、平、负三种投注的奖金分别为 100/ 90、100/ 5 和 100/ 5，当条件 1 成立时，三种投注的奖金期望值分别是：

E(投注胜的奖金)＝90％×100/ 90＝1(元)；

E(投注平的奖金)＝5％×100/ 5＝1(元)；

E(投注负的奖金)＝5％×100/ 5＝1(元)。

当条件 2 成立时，三种投注的奖金期望值分别是：

E(投注胜的奖金)＝70％×100/ 90＝0.78(元)；

E(投注平的奖金)＝15％×100/ 5＝3(元)；

E(投注负的奖金)＝15％×100/ 5＝3(元)。

通过对比两组计算结果，我们可以发现，当三种结果的投注比例与发生概率不同时，不同投注结果的奖金期望值是不同的，平均意义上讲，小概率事件由于奖金更高反而比大概率事件获得的奖金更多，这就是利用冷门提高奖金期望值的方法。在购买足彩时，我们应当留心那些可能爆冷的比赛，当你认为强队的获胜概率被高估了，就应当坚决的选择冷门结果！

我们尝试把上面的策略推广到多场比赛。

假设我们竞猜两场比赛的胜平负结果，投注共计 100 注，总奖金 100 元，投注结果如表 5-2 所示。两场比赛中强队取胜、打平和告负的概率分别为 70％、15％和 15％，此时，哪一种投注方式的奖金期望更高？

表 5-2　两场比赛的投注结果

投注	数量
胜胜	60
胜平	9
胜负	9
平胜	9
平平	1
平负	1
负胜	9
负平	1
负负	1

这 9 种投注方式的奖金期望是：

E(胜胜)=70%×70%×100/ 60=0.82(元)；

E(胜平)=70%×15%×100/ 9=1.17(元)；

E(胜负)=70%×15%×100/ 9=1.17(元)；

E(平胜)=70%×15%×100/ 9=1.17(元)；

E(平平)=15%×15%×100=2.25(元)；

E(平负)=15%×15%×100=2.25(元)；

E(负胜)=70%×15%×100/ 9=1.17(元)；

E(负平)=15%×15%×100=2.25(元)；

E(负负)=15%×15%×100=2.25(元)。

由此可见，平均意义上，投注两场都出冷门依然是获利更高的投注方式。虽然计算结果与假设条件密切相关，但不可否认的是，搏冷门并非冲动之举，是有概率原理支持的。不过，我并不鼓励投注者全力搏冷门，可以预见的是，如果我们要投注 3 场比赛，全部选择冷门结果的奖金期望会低于选择一场或两场冷门的投注。所以，搏冷门绝非多多益善。

技巧四：合理进行复式投注

足彩中的复式投注是指，同时选择一场的多个结果，然后把所有可能结果的组合一起购买。例如，下面两场比赛：

皇家马德里 VS 瓦伦西亚

巴塞罗那 VS 马德里竞技

单式投注是类似“胜胜”“胜平”的投注，复式投注则是类似“胜、胜平”“胜

负、胜平”的投注。当你投注“胜、胜平”时，意味着你购买了两注单式——“胜胜”和“胜平”，投注“胜负、胜平”则相当于投注了四注单式。

买过足彩的朋友一定会纠结一个问题：如何利用复式投注预防冷门？以上面两场比赛为例，皇家马德里和巴塞罗那在主场取胜的概率自然很高，可是我们已经知道了，搏冷门可以提高奖金期望值，那么，如果允许复式投注，我们应当怎么做呢？

以皇家马德里 VS 瓦伦西亚的比赛为例，仍然假设皇家马德里赢球的概率是 70%，打平和输球的概率各为 15%，投注总计 100 注，其中 90 注选择胜，5 注选择平，5 注选择负，总奖金依然是 100 元。此时，我们进行复式投注，同时选择两个结果，该如何选择？

选择两个结果，有三种可能的组合——“胜平”“胜负”和“平负”，分别计算三种选择的奖金期望，可以得到：

E(胜平的奖金)＝70%×100/90＋15%×100/5＝3.78(元)；

E(胜负的奖金)＝70%×100/90＋15%×100/5＝3.78(元)；

E(平负的奖金)＝15%×100/5＋15%×100/5＝6(元)。

计算结果说明，“平负”的奖金期望值最高，对于一场比赛来说，如果某支球队被高估，那么放弃热门结果，全部选择冷门结果是更合理的。这个结论同样不能简单外推至多场比赛，这与前文对冷门的讨论类似。

“14 场胜负彩”总结：

(1) 一、二等奖的中奖概率极低；

(2) 博彩公司的赔率可以作为实力对比的参考指标；

(3) 要懂球，会做基本的实力分析；

(4) 可以挑选一场或两场强弱分明的比赛，用单式或复式投注搏冷门。

5.4　得州扑克：我不是教你诈

在东北，逢年过节打麻将、打扑克是每个家庭必备的娱乐项目，我家也不例外。在我还很小的时候，麻将是我的强项，我要赢钱几乎不需要技巧，因为大人们的手气从来都比不过一个 6 岁的孩子！也正因为这个，大人们渐渐开

始不欢迎我，后来，我只好坐在炕头跟堂哥们打扑克了。我有两位堂哥，大哥大我六岁，二哥大我两岁。大哥会教我们玩很多种扑克游戏——刨幺、升级、红十，他似乎什么都会玩。这些扑克游戏虽然需要技巧，但也依靠手气，所以我依然可以靠手气赢到钱。唯独有一个游戏，我没法凭手气赢到钱，这个游戏叫作“帕斯”。

“帕斯”，是我们的口头叫法，大概是“PASS”的音译，玩法很简单。拿来一副扑克牌，去掉大小王，剩下 52 张。几个人围坐一圈，每轮每人摸一张牌，一共摸五轮，前两轮牌面向下，后三轮牌面向上，最后，比较五张牌的大小，牌最大的玩家算赢。在每轮发牌之后，按照明牌的大小顺序依次下注，后一个玩家可以选择跟注、加倍或弃牌，跟注就是与上家下注同样的赌金，加倍则表示你要比上家下注的多一倍、两倍甚至三倍，每当有玩家加倍，其他玩家必须跟注同样的赌金才能继续留在赌局中，弃牌则意味着退出游戏，输掉此前下注的所有赌金。“帕斯”最有意思的部分是“诈”，这恰恰是大哥最擅长的，也恰恰是我最不擅长的。所谓“诈”，就是“诈唬”，用加倍来诈唬对手，让对手弃牌，不战而胜。大哥是“诈唬”高手，他亦虚亦实的“诈唬”让我防不胜防，即使我可以凭手气赢下几局，也难逃输钱的结果。长大以后，我方才知道，“帕斯”是一种知名扑克游戏的变种，它就是“得州扑克”。

得克萨斯扑克，简称得州扑克，起源于 20 世纪初的美国得克萨斯州洛布斯镇，传播至赌城拉斯维加斯后，被广为传播。得州扑克是每年世界扑克大赛的主要赛事，在当下的美国非常流行，近年来，随着网络社交游戏进入我们的生活，得州扑克在中国也逐渐流行起来。得州扑克与“帕斯”非常相近的一点是“诈”，比“帕斯”更刺激的是，得州扑克是无限下注游戏，即玩家加倍下注时可以加注任意多的赌金，甚至“梭哈”——下注全部赌金。“诈”固然是一种赢钱的手法，但是仅仅依靠“诈”，你一定赢不到钱，因为“得州高手”能识破你的“诈”！接下来，我们回归理性，从概率统计的角度解读得州扑克，助你迈出“得州高手”的第一步！

游戏规则

得州扑克的规则如下所述。

台面围坐约 2～10 人，使用一副扑克牌，去掉大小王，共 52 张牌。每个玩家分两张牌，作为“底牌”，底牌牌面向下，每个玩家只知道自己的底牌。然后，开始发公共牌，公共牌牌面向上，一共五张。在底牌和每张公共牌发完后，都要进行下注，下注同样分为跟注、加倍和弃牌，加倍最低要是上一个玩家的两倍，上不封顶。所有公共牌都发完并且所有下注都完成后，所有玩家摊牌，比较大小。比较的方法是：两张底牌与五张公共牌混合后，所能选出的最“大”的五张牌就是玩家的牌面大小。

得州扑克在比较牌面大小时，首先比较牌型，牌型大的是赢家，表 5-3 列出了得州扑克的牌型大小顺序。例如，图 5-4 是 X、Y、Z 三个玩家的牌局示意图，X 的牌型是顺子（由底牌♥6、♣7 和公共牌♣9、◆8、◆5 组成），Y 的牌型是三条（由底牌♠5、♣5 和公共牌◆5、♠A、♣K 组成），Z 的牌型是一对（由底牌♣A 和公共牌♠A、♣K、♣9、◆8 组成），根据牌型的大小顺序，X 比 Y 和 Z 的牌型都更大，X 是赢家。

表 5-3　得州扑克牌型

牌　　型	示　　例	注　　释
皇家同花顺	♥A ♥K ♥Q ♥J ♥10	最大为 A 的同花顺
同花顺	♣9 ♣8 ♣7 ♣6 ♣5	花色相同的顺子
四条	♠5 ♥5 ♣5 ◆5 ♥8	四张牌点数相同
三带二	♥K ♣K ◆K ♠3 ◆3	三条和一对的组合
同花	♠2 ♠5 ♠7 ♠J ♠A	花色相同的五张牌
顺子	♠3 ♣4 ♠5 ♥6 ◆7	点数相连的五张牌
三条	◆7 ♥7 ♠7 ♣Q ♠K	三条和两张散牌
两对	♥5 ♠5 ♣9 ◆9 ♠K	两个对子
一对	◆Q ♠Q ♥3 ◆6 ♠7	一个对子和三张散牌
散牌	♥A ♠Q ♣J ◆9 ♥5	五张散牌

若牌型相同，则按照牌型从大到小比较点数，例如，“♠3、♣3、◆3、♥A、♠A”和“♠9、♣9、◆9、♥4、♠4”都是三带二的牌型，但是由于 9 点大于 3 点，所以后者比前者大，“♠A、♥J、◆5、♥4、♣3”和“◆K、♥Q、◆J、♥9、♣8”都是散牌，♠A 比◆K 大，因此前者大。

下面，我们就从玩家的角度来研究一下得州扑克中的概率原理。

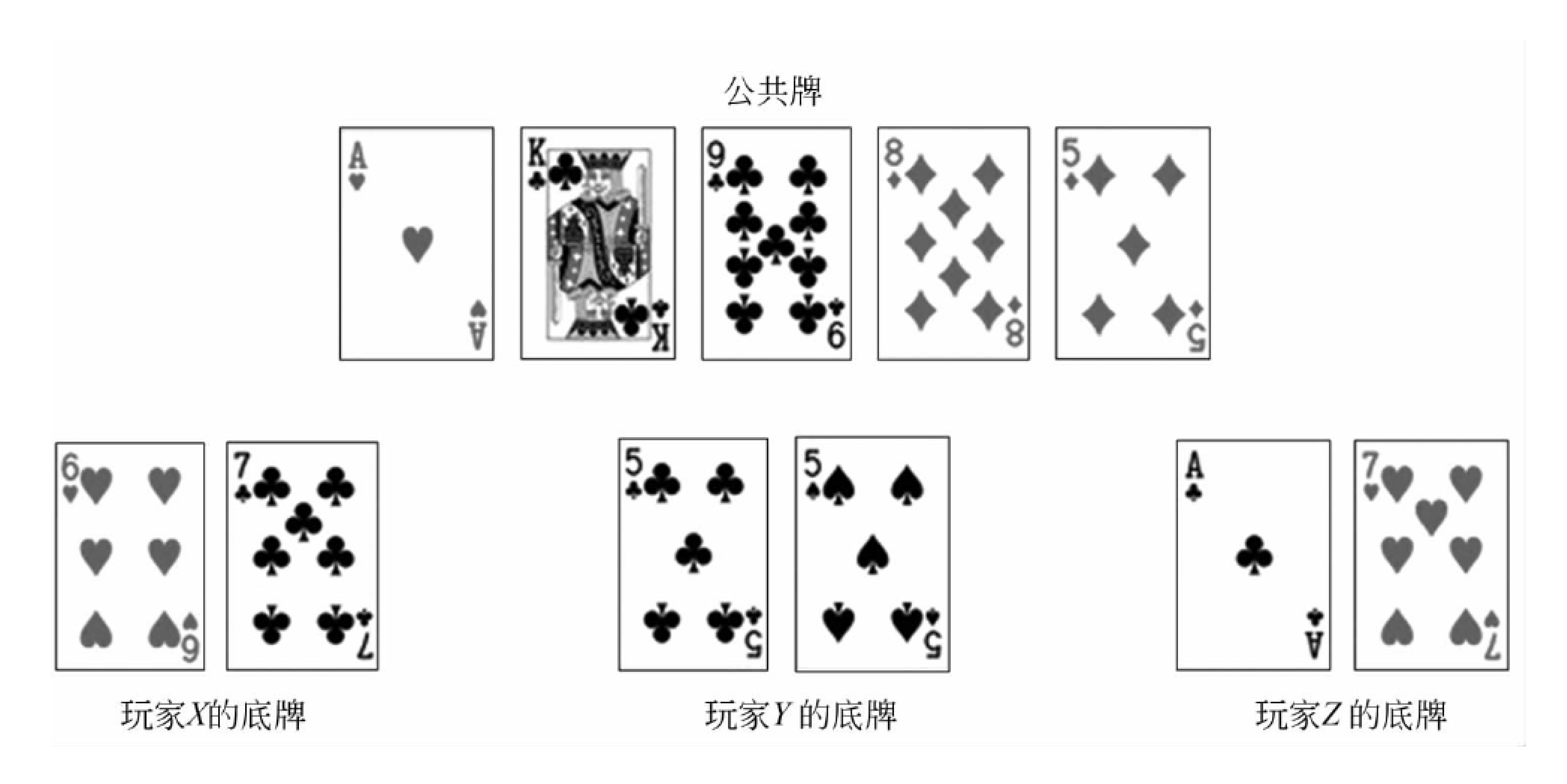


图 5-5　得州扑克牌局示意图

发牌前

在每一局发牌前，我们对可能出现的结果应该心中有数。公共牌有 5 张，再算上玩家手中的底牌，一共是 7 张，因此，我们需要计算出两张概率表：一是 5 张牌出现各种牌型的概率，如表 5-4 所示；二是 7 张牌出现各种牌型的概率，如表 5-5 所示。其计算过程涉及排列组合的知识，表 5-4 中给出了计算公式，供读者参考。

表 5-4　5 张牌出现各种牌型的概率

牌　　型	出现概率(%)	计 算 公 式
皇家同花顺	0.000 15	$P_1=4/C_{52}^5$
同花顺	0.001 4	$P_2=4\times 9/C_{52}^5$
四条	0.024	$P_3=13\times C_{48}^1/C_{52}^5$
三带二	0.14	$P_4=13\times C_4^3\times 12\times C_4^2/C_{52}^5$
同花	0.20	$P_5=(4\times C_{13}^5-4\times 10)/C_{52}^5$
顺子	0.39	$P_6=(4^5-4)/C_{52}^5$
三条	2.11	$P_7=13\times C_4^3\times C_{12}^2\times C_4^1\times C_4^1/C_{52}^5$
两对	4.75	$P_8=C_{13}^2\cdot C_4^2\cdot C_4^2\cdot C_{44}^1/C_{52}^5$
一对	42.26	$P_9=13\times C_4^2\times C_{12}^3\times C_4^1\times C_4^1\times C_4^1/C_{52}^5$
散牌	50.12	$P_{10}=1-P_1-P_2-P_3-P_4-P_5-P_6-P_7-P_8-P_9$

表 5-5　7 张牌出现各种牌型的概率

牌　　型	出现概率(%)
皇家同花顺	0.003 2
同花顺	0.028
四条	0.17
三带二	2.60
同花	3.03
顺子	4.62
三条	4.83
两对	23.50
一对	43.82
散牌	17.41

观察表 5-4 可以发现,5 张牌出现皇家同花顺、同花顺和四条的概率非常低,出现三带二、同花和顺子的概率也比较低,出现三条和两对的概率略高一点,最可能出现的是一对和散牌。这样的概率分布可以帮助我们粗略的判断公共牌可能出现哪些情况。

观察表 5-5 可以发现,7 张牌出现皇家同花顺、同花顺和四条的概率依然非常低,出现三带二、同花、顺子和三条的概率稍高,最可能出现的三种牌型依次是散牌、两对和一对。7 张牌与 5 张牌的概率分布出现了明显的不同,三带二、同花和顺子出现的概率小幅提高了,两对出现的概率大幅度提高,此外,散牌和对子出现的概率降低了。

在不考虑其他条件的情况下,我们可以利用上述的两个概率分布表预知两件事:

(1) 公共牌很可能会出现散牌或一对;

(2) 每个玩家最后的牌型很可能是散牌、两对或一对。

底牌

每个人的底牌有两张,这两张牌是你的“秘密武器”,格外关键。在得州扑克的牌局中,有些作风保守的人,拿到诸如“梅花 8,方块 2”这样的底牌时,会直接弃牌,有些人则不论底牌怎么差也不弃牌,底牌到底有多重要?我们来算一算。

从玩家的角度来看，他只能看到自己的底牌，因此，这相当于从 52 张牌中挑选 2 张牌，这 2 张牌可能出现的牌型和对应的概率如表 5-6 所示。这张表告诉我们，底牌摸到一对的概率仅有 5.88%，所以绝大多数时候底牌都不是对子；至少摸到一张 A 的概率高达 14.9%，我相信比大多数人预想的高得多；至少一张牌不小于 J 的概率高达 52.49%，因此如果你摸到的底牌全都比 10 小且不是对子，那么你的手牌很可能不是牌局中最大的，换言之，除非公共牌对你很有利，否则你应该谨慎下注。

表 5-6　底牌的牌型和出现概率

牌　　型	出现概率(%)
一对	5.88
非一对	94.12
特定数字的一对 (例如♠K◆K、♥5 ♣5)	0.45
某个非对子牌型 (例如♠K◆9、♥5 ♣4)	1.20
至少一张 A (例如♠A◆9、♥A ♣A)	14.9
至少一张不小于 K (例如♠K◆9、♥A ♣5)	28.66
至少一张不小于 Q (例如♠K◆9、♥Q ♣5)	41.18
至少一张不小于 J (例如♥J ♣6、Q5 ♣2)	52.49

公共牌

在拿到底牌后，每个玩家都对公共牌充满期待。可是，你知道你期待的牌出现的概率是多少吗？

比如，玩家的底牌是“黑桃 8，红桃 8”，这牌还算不错，但玩家还希望公共牌中再出现至少一个 8，这样一来获胜的概率会高很多。要精确计算“公共牌中至少出现一个 8”的概率很困难，因为我们站在玩家的视角，无法看到其他玩家的底牌，所以我们只能估算：

$$P(\text{公共牌中至少出现一个 8}) = (C_{50}^{5} - C_{48}^{5}) / C_{50}^{5} = 19.2\%$$

这说明，在公共牌没有发出之前，公共牌中至少出现一个 8 的概率有 19%，这个概率并不算低。

又如，玩家的底牌是“黑桃 8，红桃 7”，已经发出的四张公共牌是“红桃 K，黑桃 5，方块 9，梅花 Q”，玩家唯一的希望就是最后一张公共牌是 6，这样就会形成顺子。我们估算一下最后一张牌出现 6 的概率：

$$P(\text{最后一张公共牌出现 6}) = 4/46 = 8.70\%$$

多么让人沮丧的结果，但这就是现实——不要对某一张牌抱太大的希望。

上面的计算还有一种简便方法：牌堆中还有 4 张 6，我们把 4 乘以 2，再加 1，得到 9，因此，6 出现的概率大约是 9%，与 8.70%的计算结果很接近。这个简便算法的原理很简单，由于扑克牌的数量大约为 50 张，因此，你想要的那张牌在牌堆中的数量乘以 2，便得到了那张牌出现的概率，再加 1 是对这一概率进行的修正，因为牌堆大多数时候都不足 50 张。这个简便算法可以帮助我们快速估算我们期待的牌在下一张出现的概率。

摊牌

当五张公共牌都发出后，牌局就进入了最刺激的末轮下注。此时，玩家需要估算自己取胜的概率，以决定如何下注。既然公共牌最可能出现的牌型是散牌，我们就以散牌为例，来计算玩家取胜的概率。

如图 5-6 所示，五张公共牌是“♥K，♠J，♠8，♠7，♦3”，玩家的底牌是“♥8，♣A”，这时，玩家首先要想到，其他玩家可能的牌型有黑桃同花顺、黑桃同花、顺子、三条、两对、一对和散牌，然后玩家需要一一估算出这些牌型出现的概率。

要出现黑桃同花顺，需要底牌是♠9 和♠10，因此：

$$P(\text{黑桃同花顺}) = 1/C_{45}^{2} = 0.10\%$$

要出现黑桃同花，需要底牌是两张黑桃牌，因此：

$$P(\text{黑桃同花}) = (C_{10}^{2} - 1)/C_{45}^{2} = 4.44\%$$

要出现顺子，需要底牌是 9 和 10，因此：

$$P(\text{顺子}) = (C_{4}^{1} \times C_{4}^{1} - 1)/C_{45}^{2} = 1.52\%$$

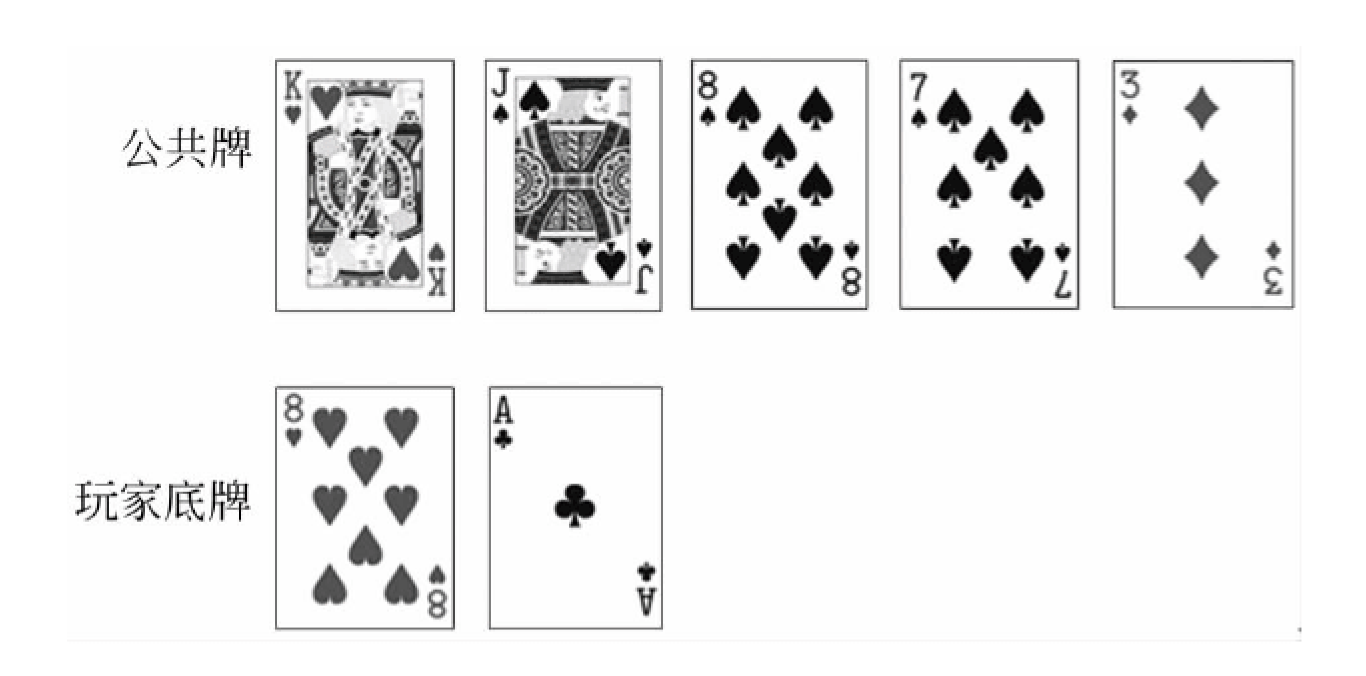


图 5-6　牌局示例

要出现三条，需要底牌是 K、J、8、7 或 3 的对子，因此：

$$P(\text{三条}) = (C_3^2 + C_3^2 + C_2^2 + C_3^2 + C_3^2)/\ C_{45}^2 = 1.31\%$$

要出现两对，需要底牌是 K、J、8、7、3 中的任意两个，因此：

$$P(\text{两对}) = (C_5^2 \cdot C_3^2 \cdot C_3^2 - C_4^1 \cdot C_3^1)/\ C_{45}^2 = 7.88\%$$

要出现一对，需要底牌中的一张是 K、J、8、7、3 中的一个，另一个不是，因此：

$$P(\text{一对}) = C_{14}^1 \cdot C_{31}^1/\ C_{45}^2 = 43.8\%$$

出现散牌的概率是：

$$\begin{aligned} P(\text{散牌}) &= 1 - P(\text{黑桃同花顺}) - P(\text{黑桃同花}) - P(\text{顺子}) - \\ &\quad P(\text{三条}) - P(\text{两对}) - P(\text{一对}) \\ &= 40.9\% \end{aligned}$$

上面的估算结果说明，其他玩家最可能的牌型是一对和散牌，其次可能的牌型是两对和黑桃同花，其他牌型出现的概率非常低。本例与表 5-5 不同之处是，黑桃同花出现的概率高于顺子和三条，这说明，玩家要根据公共牌的情况重新估算各种牌型出现的概率，不能生搬硬套表 5-5 中的概率分布。

最后，玩家还需要知道最重要的一件事是，他赢得牌局的概率有多大？

玩家有一对 8，很明显，黑桃同花顺、黑桃同花、顺子、三条和两对都比玩家的牌型大，散牌则比玩家的牌型小，玩家需要更详细的估算一对出现的概率：

$$P(\text{一对 K}) = C_3^1 \cdot C_{31}^1/\ C_{45}^2 = 9.40\%$$

$$P(\text{一对 J}) = C_3^1 \cdot C_{31}^1/\ C_{45}^2 = 9.40\%$$

$$P(\text{一对 8}) = C_2^1 \cdot C_{31}^1/\ C_{45}^2 = 6.26\%$$

$$P(\text{一对 7}) = C_3^1 \cdot C_{31}^1/\ C_{45}^2 = 9.40\%$$

$$P(\text{一对 }3) = C_3^1 \cdot C_{31}^1 / C_{45}^2 = 9.40\%$$

当其他玩家也有一对 8 的牌型时，由于玩家的另一张底牌是 A，所以玩家一定不会输，我们把这种情况也视为玩家赢，由此可以估算出玩家赢的概率：

$$P(\text{玩家赢}) = P(\text{一对 }8) + P(\text{一对 }7) + P(\text{一对 }3) + P(\text{散牌}) = 65.96\%$$

可见，虽然一对 8 并不算是大牌，但足以让玩家赢的概率达到 65.96%，这真是一个赢的好机会！

得州扑克总结：

(1) 皇家同花顺、同花顺和四条很难出现；

(2) 公共牌很可能会出现散牌或一对；

(3) 每个玩家最后的牌型很可能是一对、两对或散牌；

(4) 散牌赢的概率一般低于 50%；

(5) 不同的公共牌会有不同的概率分布，要具体问题具体分析。

提醒读者：上面的所有概率计算都是玩家的估算，只能作为下注的参考，要成为真正的得州高手，只有一个办法——现在就玩一局吧！

5.5　21 点：保守未必是坏事

“21 点”也是赌场里十分流行的扑克游戏，虽然与大转盘、老虎机一样是与庄家对决，但是 21 点和那些注定输钱的游戏有本质的不同——玩家可以自由选择策略，但庄家不能。在这个看似有利的规则下，玩家有可能从庄家手中赢到钱吗？这一节，我们就从概率统计的角度算一算 21 点玩家赢钱的概率。

游戏规则

21 点的游戏规则如下所述。

发牌者一张接一张的给玩家发牌，玩家每得到一张牌，就要计算一下手上所有牌的点数之和，然后选择继续发牌或者停止发牌。在玩家选择停止发牌后，发牌者给庄家发牌，直到庄家喊停为止。最后，双方摊牌比较大小。如果双方的总点数都不大于 21 点，则点数大的一方获胜，点数相同算打平；手牌的

总点数超过 21 点，称为“爆点”，如果玩家爆点，则直接输掉赌局，无须给庄家发牌，如果玩家没爆点，庄家爆点，则玩家赢得赌局。此外，有一种特殊牌型是一张 A 和一张 10，称为“黑杰克”，如果玩家的牌是“黑杰克”而庄家不是，玩家赢得 1.5 倍赌金；反之，庄家赢得 1.5 倍赌金，如果双方都是“黑杰克”，算打平。

21 点最重要的规则是停牌规则，玩家有权在拿到任何一张牌后停牌，但是庄家在总点数达到 17 点或 17 点以上时，必须停牌。

点数大小的计算规则是，A 是 1 点或 11 点（黑杰克牌型），K、Q、J、10 均计 10 点，其余的牌按照牌面数字计点数。这里需要说明的是，为了保证公平，21 点游戏一般会使用 6 副甚至更多副牌，这样可以保证双方每一轮得到不同点数牌的概率几乎相同。根据点数计算规则，在不考虑黑杰克牌型的前提下，双方每一轮拿到 1～9 点中某一个点数的概率是 1/13，拿到 10 点的概率是 4/13。

规则告诉我们，玩家如果爆点会直接输掉赌局，因此玩家需要理性看待爆点。根据各点数出现的概率，可以计算出下一张牌爆点的概率，如表 5-7 所示。从手牌总点数 12 点开始，爆点的概率逐渐上升，点数为 12 时，爆点概率为 30.8%；点数达到 15 时，爆点概率超过 50%；点数达到 17 时，爆点概率达到约 70%。17 点是庄家给自己设置的强制停牌点数，从表 5-7 可以看出，庄家给自己留出了约 30%的容错空间。那么，玩家应该选择怎样的策略呢？

表 5-7 爆点概率

手牌总点数	下一轮可能导致爆点的牌	下一轮爆点的概率（%）
1～11	无	0
12	10，J，Q，K	30.8
13	9，10，J，Q，K	38.5
14	8，9，10，J，Q，K	46.2
15	7，8，9，10，J，Q，K	53.9
16	6，7，8，9，10，J，Q，K	61.6
17	5，6，7，8，9，10，J，Q，K	69.3
18	4，5，6，7，8，9，10，J，Q，K	77.0
19	3，4，5，6，7，8，9，10，J，Q，K	84.7
20	2，3，4，5，6，7，8，9，10，J，Q，K	92.4
21	所有牌	100

三种策略

21 点的停牌规则给了玩家很大的自由度，玩家可以自由安排策略，这便是 21 点考验玩家智慧的地方。玩家既可以比庄家更保守，也可以比庄家更激进，还可以"以牙还牙"，采用和庄家相同的策略。三种具有代表性的策略如下所述。

（1）保守策略。玩家在手牌点数大于 11 点时选择停止发牌，保证绝不会爆点。

（2）对等策略。玩家在手牌点数大于或等于 17 点时选择停止发牌，与庄家采用同样的策略。

（3）激进策略。玩家在手牌点数大于 20 点时选择停止发牌，不得到 21 点誓不罢休。

我借助一点编程技巧完成了三种策略的概率计算，得到表 5-8。

表 5-8　三种策略的点数概率分布

点数	概　　率		
	保守策略(%)	对等策略(%)	激进策略(%)
12	12.10	—	—
13	11.61	—	—
14	11.08	—	—
15	10.52	—	—
16	9.90	—	—
17	9.24	14.23	—
18	8.53	13.52	—
19	7.77	12.76	—
20	12.27	17.26	—
21	2.24	7.22	12.18
黑杰克	4.73	4.73	4.73
爆点	0.00	30.28	83.09

利用表中的概率分布可以进一步计算出三种策略不同点数的赢、平、输条件概率，如表 5-9～表 5-11 所示。

表 5-9　保守策略的赢平输概率分布

点数	赢的概率(%)	平的概率(%)	输的概率(%)
12	30.28	0.00	69.72
13	30.28	0.00	69.72
14	30.28	0.00	69.72
15	30.28	0.00	69.72
16	30.28	0.00	69.72
17	30.28	14.23	55.49
18	44.51	13.52	41.97
19	58.03	12.76	29.21
20	70.79	17.26	11.95
21	88.05	7.22	4.73
黑杰克	95.27	4.73	0.00
爆点	0.00	0.00	100.00

表 5-10　对等策略的赢平输概率分布

点数	赢的概率(%)	平的概率(%)	输的概率(%)
17	30.28	14.23	55.49
18	44.51	13.52	41.97
19	58.03	12.76	29.21
20	70.79	17.26	11.95
21	88.05	7.22	4.73
黑杰克	95.27	4.73	0.00
爆点	0.00	0.00	100

表 5-11　激进策略的赢平输概率分布

点数	赢的概率(%)	平的概率(%)	输的概率(%)
21	88.05	7.22	4.73
黑杰克	95.27	4.73	0.00
爆点	0.00	0.00	100.00

借助表 5-9～表 5-11 的概率分布，可以计算出三种策略的赢平输概率，如表 5-12 所示。表 5-12 中数据显示，保守策略和对等策略都有较高的胜率，输的概率都接近 50%，但激进策略的表现则很糟糕，输的概率高达 83%。因此，要想在21 点游戏中争取更多胜利，宁可保守也不可冒进。

表 5-12　三种策略的赢平输概率

	赢的概率(%)	平的概率(%)	输的概率(%)
保守策略	42.99	5.96	51.05
对等策略	40.81	9.21	49.98
激进策略	15.23	1.10	83.67

只计算输赢的概率还不够,收益的期望值才能真正反映策略的优劣。假设玩家和庄家的赌金都是 100 元,黑杰克出现时会赢得 150 元的赌金,由此得到三种策略的收益期望分别是:

$$E(\text{保守策略的收益}) = -8.06(\text{元})$$

$$E(\text{对等策略的收益}) = -8.45(\text{元})$$

$$E(\text{激进策略的收益}) = -66.47(\text{元})$$

保守策略的收益期望值依然是最高的,其次是对等策略,激进策略依然是最糟糕的选择,无论三者孰高孰低,三种策略的收益期望都是负数,根据大数定理,连续不断地玩下去,玩家一定会输钱,不同的策略只是影响输钱的快慢和多少罢了。

此外,有一个问题不知读者有没有想过:既然玩家采取了和庄家一样的对等策略,为什么收益期望值还是负数呢?难道双方不应该打平吗?答案是"爆点"的规则打破了玩家和庄家的平衡,当玩家爆点时,会直接输掉赌金,如此一来,庄家就没有机会"爆点"了,收益的天平就向庄家倾斜了。

最后,我们可以用一句话总结 21 点游戏:保守一些总不会错!

第 6 章

假 设 检 验

导语：主场作战意味着熟悉的更衣室、熟悉的地板、熟悉的篮筐和球迷们山呼海啸的助威，所以主场作战的球队总是会获胜。体育世界中所谓的“主场优势”，是媒体的造势还是确有其事？假设检验为你揭开谜底。

6.1 主场优势：规律还是假象？

“中国奥运代表团在2008年北京奥运会上实现了突破，首次获得金牌榜第一名！”

“利物浦队坐镇安菲尔德球场三球大胜来访的曼联队！”

“比赛结束了！勇士队在自己的主场输给了凯尔特人队，终结了主场54连胜的纪录！”

主场是体育迷最熟悉的一个词，主场作战意味着熟悉的更衣室、熟悉的地板、熟悉的篮筐，进球时可以接受全场球迷的喝彩，落后时会听到山呼海啸的加油！每场比赛前，主场作战的球队都会受到媒体和球迷的偏爱，只因为每个

人都知道，主场作战的球队握有独一无二的武器——主场优势。

主场优势是体育世界里的一个自然形成的“规律”，虽然主场球队和客场球队在同样的天气下、同样的场地上比赛，但是主场球队似乎总是表现得更好。主场优势到底是媒体的造势还是确有其事？我们从球迷们最熟悉的两项赛事说起。

NBA的主场优势

北美职业篮球联盟（National Basketball Association，NBA）代表了篮球运动的最高水平，新赛季从每年的 10 月持续至次年 6 月，联盟的 30 支球队分东西两个半区进行比赛。比赛分为循环赛和淘汰赛两个阶段，循环赛称为常规赛，每支球队都要打满 82 场常规赛。接下来，东西半区排名前八名的球队进入 7 场 4 胜的淘汰赛——季后赛。季后赛的竞争总是异常激烈，充斥着强悍的身体对抗、地板球争抢甚至粗暴的犯规，火药味儿十足。主场优势在季后赛中也得以彰显，主队的每一个进球、每一次成功防守都会引发全场球迷的喝彩，客队球员每一次罚球都会遭到球迷们肆无忌惮的干扰。在这样的氛围下，客队很难带走一场胜利。

我们用数据说话。2014－2015 赛季的 NBA 常规赛一共进行了 1 230 场比赛，主队取胜 707 场，胜率 57.5％；2014－2015 赛季的季后赛一共进行了 81 场比赛，主队取胜 48 场，胜率达到 59.3％。表 6-1 是 2014－2015 赛季 NBA 常规赛部分赛果，表 6-2 是 2014－2015 赛季 NBA 季后赛部分赛果，数据来自美国体育数据网站 Sports Reference（网站地址：http://www.basketball-reference.com）。

表 6-1　2014—2015 赛季 NBA 常规赛部分赛果

场次	客　　队	客队得分（分）	主　　队	主队得分（分）	主队赛结果
1	休斯敦火箭	108	洛杉矶湖人	90	负
2	奥兰多魔术	84	新奥尔良鹈鹕	101	胜
3	达拉斯小牛	100	圣安东尼奥马刺	101	胜
4	布鲁克林篮网	105	波士顿凯尔特人	121	胜
5	密尔沃基雄鹿	106	夏洛特黄蜂	108	胜

续表

场次	客　　队	客队得分（分）	主　　队	主队得分（分）	主队赛结果
6	底特律活塞	79	丹佛掘金	89	胜
7	费城 76 人	91	印第安纳步行者	103	胜
8	明尼苏达森林狼	101	孟菲斯灰熊	105	胜
9	华盛顿奇才	95	迈阿密热火	107	胜
10	芝加哥公牛	104	纽约尼克斯	80	负

表 6-2　2014—2015 赛季 NBA 季后赛部分赛果

场次	客　　队	客队得分（分）	主　　队	主队得分（分）	主队赛结果
1	密尔沃基雄鹿	91	芝加哥公牛	103	胜
2	新奥尔良鹈鹕	99	金州勇士	106	胜
3	达拉斯小牛	108	休斯敦火箭	118	胜
4	华盛顿奇才	93	多伦多猛龙	86	负
5	布鲁克林篮网	92	亚特兰大老鹰	99	胜
6	波士顿凯尔特人	100	克里夫兰骑士	113	胜
7	圣安东尼奥马刺	92	洛杉矶快船	107	胜
8	波特兰开拓者	86	孟菲斯灰熊	100	胜
9	密尔沃基雄鹿	82	芝加哥公牛	91	胜
10	新奥尔良鹈鹕	87	金州勇士	97	胜

这些数据似乎从统计意义上说明了，主队的确更容易获胜。可是，我们依然可以找到反例，比如，2014－2015 赛季的总决赛，勇士队和骑士队一共进行了 6 场比赛，主队和客队各取胜 3 场。又如，2013－2014 赛季的总决赛，马刺队和热火队一共进行了 5 场比赛，主队只取胜 2 场，客队却取胜了 3 场。这些反例在提醒我们，主场优势并非时刻都会显现，经验老到的马刺队和三分无解的勇士队都曾反客为主，逆势取胜。

世界杯的主场优势

2014 年 6 月 12 日，第 20 届世界杯在热辣的桑巴舞曲中拉开大幕。“桑巴军团”巴西队坐镇主场，气势如虹，“潘帕斯雄鹰”阿根廷队同为南美老乡，也算拥有半个主场，再加上队中有梅西、阿奎罗等一流攻击手，也志在夺冠。两支主场作战的球队不负众望，一起杀进了半决赛。

半决赛第一场,巴西队遭遇德国队。比赛进行了不到 30 分钟,场边的数万巴西球迷便已心碎,比赛俨然成为德国队的进攻表演,90 分钟过后,比分牌上赫然显示着 7∶1,巴西队在家乡父老面前刷新了一个耻辱的记录——世界杯半决赛的最大分差。半决赛第二场,阿根廷队迎战老对手荷兰队,双方鏖战到加时赛时依然难分高下,点球大战中,阿根廷门将罗梅罗扑出了荷兰队的两粒点球,力助阿根廷队挺进决赛。

决赛前,全球媒体对比赛结果做出了很多预测,论实力,德国队略胜一筹,但在一场定胜负的决赛中,以弱胜强的案例数不胜数,而且有一项统计数据给德国队夺冠蒙上了阴影——美洲举办的世界杯上夺冠的都是美洲球队。

翻开世界杯的史册,在 2014 年巴西世界杯之前,共有 7 届世界杯赛在美洲国家举办,最终捧杯的都是美洲球队,而且其中有 5 次是美洲球队战胜欧洲球队夺冠。更不利于德国队的是,半决赛上对巴西队的羞辱"激怒"了巴西甚至全南美洲的球迷,他们纷纷穿上阿根廷的球衣,把决赛场地彻底变成了阿根廷队的主场。决赛的过程也正如媒体所料,阿根廷队并未落入下风,梅西甚至获得过终结比赛的机会。然而,阿根廷队在加时赛中的一次防守松懈彻底葬送了比赛,他们以一球惜败,饮恨决赛。也许,只有高傲顽强的日耳曼战车才能碾碎"美洲无冠"的魔咒!

表 6-3　历届世界杯冠军

届　数	举办年份(年)	举办国	冠　军	决 赛 比 分
第一届	1930	乌拉圭	乌拉圭队	乌拉圭 4∶2 阿根廷
第二届	1934	意大利	意大利队	意大利 2∶1 捷克斯洛伐克
第三届	1938	法国	意大利队	意大利 4∶2 匈牙利
第四届	1950	巴西	乌拉圭队	乌拉圭 2∶1 巴西
第五届	1954	瑞士	西德队	西德 3∶2 匈牙利
第六届	1958	瑞典	巴西队	巴西 5∶2 瑞典
第七届	1962	智利	巴西队	巴西 3∶1 捷克斯洛伐克
第八届	1966	英格兰	英格兰队	英格兰 4∶2 西德
第九届	1970	墨西哥	巴西队	巴西 4∶1 意大利
第十届	1974	西德	西德队	西德 2∶1 荷兰
第十一届	1978	阿根廷	阿根廷队	阿根廷 3∶1 荷兰
第十二届	1982	西班牙	意大利队	意大利 3∶1 西德

续表

届　数	举办年份（年）	举办国	冠　军	决 赛 比 分
第十三届	1986	墨西哥	阿根廷队	阿根廷 3∶2 西德
第十四届	1990	意大利	西德队	西德 1∶0 阿根廷
第十五届	1994	美国	巴西队	巴西 4∶3 意大利（点球）
第十六届	1998	法国	法国队	法国 3∶0 巴西
第十七届	2002	韩国、日本	巴西队	巴西 2∶0 德国
第十八届	2006	德国	意大利队	意大利 5∶3 法国（点球）
第十九届	2010	南非	西班牙队	西班牙 1∶0 荷兰
第二十届	2014	巴西	德国队	德国 1∶0 阿根廷

不论是 NBA 还是世界杯，不论是篮球还是足球，主场优势总会成为大家热议的话题，很多统计数据都可以说明主场优势的存在，也有很多球队能够逆势取胜，我们到底应该怎样看待主场优势呢？接下来，我们就用“假设检验”来回答这个问题。

6.2　假设检验：主场真的有优势吗？

假设检验是统计推断的一种常用方法，简言之就是“先假设、再检验”。例如，在庞加莱与面包的故事中（参见“正态分布”一节），庞加莱知道面包的重量服从正态分布，但是不知道正态分布的参数，这时，庞加莱假设面包重量的均值为某个常数，利用记录的称重数据验证假设是否成立。

下面，我们就用假设检验的方法来验证主场优势是否真的存在。

定义主场优势

要验证主场是否有优势，首先要从概率统计的角度来定义主场优势。在一个 NBA 赛季中，每支球队会进行 82 场常规赛，主客场各 41 场，在常规赛结束时会得到一张战绩表，表 6-4 是 2014－2015 赛季 NBA 常规赛战绩表，表中列出了 30 支球队的总战绩和主客场战绩。为了用一个数字表现出各支球队

的战绩优劣，NBA 联盟会计算出各支球队的胜率——胜场数/总场次。表 6-4 中列出了 30 支球队的总胜率和主客场胜率。所谓主场优势，是指一支球队的主场表现优于客场表现，因此，我们用胜率差——主场胜率和客场胜率的差值——来度量一支球队主客场表现的差距。

表 6-4 列出了 30 支球队的胜率差。我们观察到，只有篮网队的胜率差是 0，其余球队的胜率差都大于 0，开拓者队的胜率差更是超过了 30%。单凭观察，我们几乎可以断定主场优势是普遍存在的，但是这还不够，要从概率统计的角度证明主场优势存在，就需要使用假设检验。

表 6-4　2014—2015 赛季 NBA 常规赛战绩表

	排名	球队	总战绩	主场战绩	客场战绩	总胜率(%)	主场胜率(%)	客场胜率(%)	胜率差(%)
东部赛区排名	1	老鹰	60 胜 22 负	35 胜 6 负	25 胜 16 负	73.20	85.37	60.98	24.39
	2	骑士	53 胜 29 负	31 胜 10 负	22 胜 19 负	64.60	75.61	53.66	21.95
	3	公牛	50 胜 32 负	27 胜 14 负	23 胜 18 负	61.00	65.85	56.10	9.76
	4	猛龙	49 胜 33 负	27 胜 14 负	22 胜 19 负	59.80	65.85	53.66	12.20
	5	奇才	46 胜 36 负	29 胜 12 负	17 胜 24 负	56.10	70.73	41.46	29.27
	6	雄鹿	41 胜 41 负	23 胜 18 负	18 胜 23 负	50.00	56.10	43.90	12.20
	7	凯尔特人	40 胜 42 负	21 胜 20 负	19 胜 22 负	48.80	51.22	46.34	4.88
	8	篮网	38 胜 44 负	19 胜 22 负	19 胜 22 负	46.30	46.34	46.34	0.00
	9	步行者	38 胜 44 负	23 胜 18 负	15 胜 26 负	46.30	56.10	36.59	19.51
	10	热火	37 胜 45 负	20 胜 21 负	17 胜 24 负	45.10	48.78	41.46	7.32
	11	黄蜂	33 胜 49 负	19 胜 22 负	14 胜 27 负	40.20	46.34	34.15	12.20
	12	活塞	32 胜 50 负	18 胜 23 负	14 胜 27 负	39.00	43.90	34.15	9.76
	13	魔术	25 胜 57 负	13 胜 28 负	12 胜 29 负	30.50	31.71	29.27	2.44
	14	76 人	18 胜 64 负	12 胜 29 负	6 胜 35 负	22.00	29.27	14.63	14.63
	15	尼克斯	17 胜 65 负	10 胜 31 负	7 胜 34 负	20.70	31.71	17.07	14.63
西部赛区排名	1	勇士	67 胜 15 负	39 胜 2 负	28 胜 13 负	81.70	95.12	68.29	26.83
	2	火箭	56 胜 26 负	30 胜 11 负	26 胜 15 负	68.30	73.17	63.41	9.76
	3	快船	56 胜 26 负	30 胜 11 负	26 胜 15 负	68.30	73.17	63.41	9.76
	4	开拓者	51 胜 31 负	32 胜 9 负	19 胜 22 负	62.20	78.05	46.34	31.71
	5	灰熊	55 胜 27 负	31 胜 10 负	24 胜 17 负	67.10	75.61	58.54	17.07

续表

	排名	球　队	总战绩	主场战绩	客场战绩	总胜率（%）	主场胜率（%）	客场胜率（%）	胜率差（%）
西部赛区排名	6	马刺	55 胜 27 负	33 胜 8 负	22 胜 19 负	67.10	80.49	53.66	26.83
	7	小牛	50 胜 32 负	27 胜 14 负	23 胜 18 负	61.00	65.85	56.10	9.76
	8	鹈鹕	45 胜 37 负	28 胜 13 负	17 胜 24 负	54.90	68.29	41.46	26.83
	9	雷霆	45 胜 37 负	29 胜 12 负	16 胜 25 负	54.90	70.73	39.02	31.71
	10	太阳	39 胜 43 负	22 胜 19 负	17 胜 24 负	47.60	53.66	41.46	12.20
	11	爵士	38 胜 44 负	21 胜 20 负	17 胜 24 负	46.30	51.22	41.46	9.76
	12	掘金	30 胜 52 负	19 胜 22 负	11 胜 30 负	36.60	46.34	26.83	19.51
	13	国王	29 胜 53 负	18 胜 23 负	11 胜 30 负	35.40	43.90	26.83	17.07
	14	湖人	21 胜 61 负	12 胜 29 负	9 胜 32 负	25.60	29.27	21.95	7.32
	15	森林狼	16 胜 66 负	9 胜 32 负	7 胜 34 负	19.50	21.95	17.07	4.88

双边 Z检验

双边 Z 检验是假设检验中的一种检验方法，我们首先学习双边 Z 检验的原理，再利用双边 Z 检验来验证主场优势。

假定主客场胜率差 X 服从正态分布 $N(\mu,\sigma_0^2)$，σ_0是已知的常数，μ 是未知参数。构造如下的两个对立的假设：

原假设 H_0：$\mu=\mu_0$

备择假设 H_1：$\mu\neq\mu_0$

原假设 H_0表示胜率差的均值是μ_0，备择假设 H_1表示胜率差的均值不是μ_0。假设检验的思路是，假设 H_0 成立，并由 H_0得到的若干推论，如果这些推论与已知条件矛盾，说明 H_0不成立，反之 H_0成立。

在 H_0成立的前提下，胜率差 X 服从均值为μ_0，标准差为σ_0的正态分布，即 $X\sim N(\mu,\sigma_0^2)$。我们知道，在采样数量足够（一般不少于 30 个）的前提下，采样数据的均值 $\overline{X}$ 应该是随机变量 X 的均值μ 的无偏估计，即，$\overline{X}$ 应该能够反映μ 的大小。因此，$\overline{X}$ 与μ 的偏差$|\overline{X}-\mu|$应该不会太大。在概率统计中，“应该不会”意味着发生的概率很小，这个“小概率”在假设检验中称为显著性水平，记

为 α，一般取值为0.05 或 0.01，即，当 H_0 成立时，$|\bar{X}-\mu|$ 非常大的概率不超过 α。

由 $X\sim N(\mu,\sigma^2)$ 可知，$Z=(\bar{X}-\mu_0)/(\sigma/\sqrt{n})$ 服从 $N(0,1)$ 标准正态分布，因此“$|\bar{X}-\mu|$ 非常大的概率不超过 α”等价于“Z 非常大的概率不超过 α”。根据正态分布的定义可以找到 Z 的取值区间，图 6-1 中的阴影部分是使得“Z 非常大的概率超过 α”的取值区间，称为拒绝域，当的值落在拒绝域中时，说明 $\bar{X}$ 与 μ 的偏差过大，我们不接受 H_0 的假设，当 Z 的值落在拒绝域之外时，说明 $\bar{X}$ 与 μ 的偏差不大，我们接受 H_0 的假设。

要确定拒绝域的位置，只需要计算出两个临界点 $-Z_{\alpha/2}$ 和 $Z_{\alpha/2}$，$-Z_{\alpha/2}$ 和 $Z_{\alpha/2}$ 是标准正态分布的 $-\alpha/2$ 分位点和 $\alpha/2$ 分位点。标准正态分布的 $\alpha/2$ 分位点 $Z_{\alpha/2}$ 是指，标准正态分布概率密度曲线上满足 $P(X>x_0)=\alpha/2$ 的 x_0 值，一般记为 $Z_{\alpha/2}$，其他分位点的定理与此类似。标准正态分布的分位点不需要计算，查“标准正态分布表”便可以得到。在双边 Z 检验中之所以将 X 变换为标准正态分布，就是为了便于查找分位点，这样的“标准化变换”是求解数学问题的常用方法。

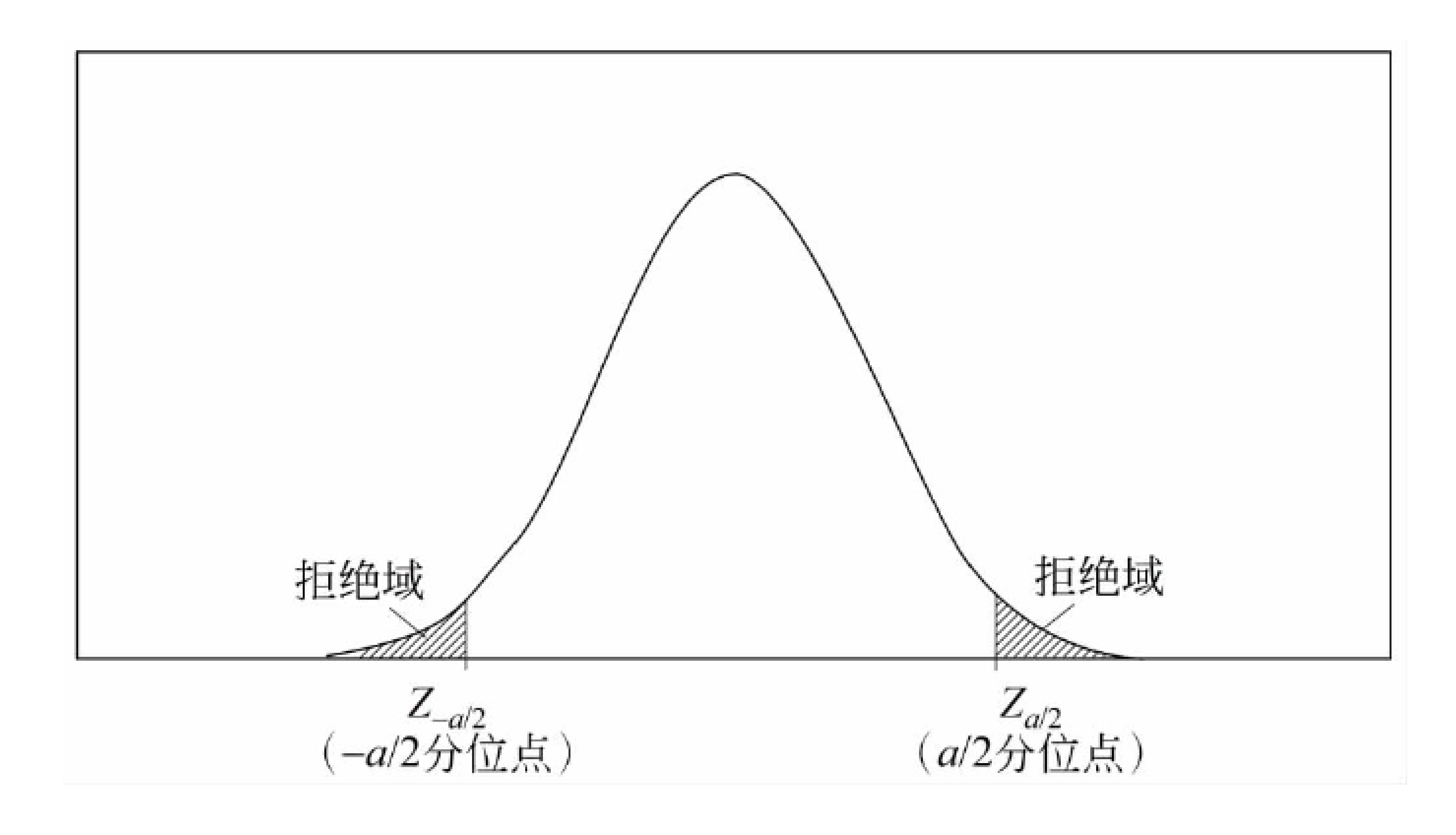


图 6-1　双边 Z 检验的临界点和拒绝域

以上便是双边 Z 检验的原理，接下来，我们便利用双边 Z 检验来验证主场优势。

主场优势可以用主客场胜率差的均值来度量，如果我们能够说明主客场胜率差的均值为某个正数，就可以说明主场优势的确存在。观察表 6-4 中的数据可以发现，有 10 支球队的胜率差十分接近 10%，因此，我们不妨令 $\mu_0=$

0.1。假设我们已知胜率差 X 服从正态分布 $N(\mu,\sigma^2)$，标准差 σ 为 0.9(由 30 支球队战绩估算的总体标准差)，即 $X\sim N(\mu,0.9^2)$。

构造如下两个假设：

H_0：$\mu=0.1$

H_1：$\mu\neq0.1$

假设 H_0 成立，则 $X\sim N(0.1,0.9^2)$。表 6-4 中的胜率差一列是 X 的 30 个采样数据，$Z=(\overline{X}-0.1)/(0.9/\sqrt{30})$服从 $N(0,1)$标准正态分布，采样数据的均值为 0.15，对应的 Z 值为$(0.15-0.1)/(0.9/\sqrt{30})$，即 0.30。

取显著性水平 $\alpha=0.05$，对应的临界点为 -1.96 和 1.96。由于 $-1.96<0.30<1.96$，因此采样数据的均值没有落入拒绝域，因此我们接受 H_0 假设，即“主客场胜率差的均值为 10%”是正确的。

如果我们把μ_0设为较大的值，则会使得均值落入拒绝域中。例如，构造如下两个假设：

H_0：$\mu=0.5$

H_1：$\mu\neq0.5$

假设 H_0 成立，则 $X\sim N(0.5,0.9^2)$，$Z=(\overline{X}-0.5)/(0.9/\sqrt{30})$服从 $N(0,1)$ 标准正态分布，采样数据的均值为 0.15，对应的 Z 值为$(0.15-0.5)/(0.9/\sqrt{30})$，即 -2.13。

取显著性水平 $\alpha=0.05$，对应的临界点为 -1.96 和 1.96。由于 $-2.13<-1.96$，因此采样数据的均值落入拒绝域，因此我们拒绝 H_0 假设，即“主客场胜率差的均值为 50%”是不正确的。

单边 Z 检验

要验证胜率差的均值大于某个常数，就需要使用单边 Z 检验。同样假定我们已知主客场胜率差 X 服从正态分布 $N(\mu,\sigma_0^2)$。构造如下的两个对立的假设：

H_0：$\mu\geqslant0.1$

H_1：$\mu<0.1$

因为 H_0 中的 μ 都比 H_1 中的大，当 H_1 为真时，样本均值会偏小，因此，拒绝域的形式为：

$$\overline{X} < k(k\text{ 是某个常数})$$

我们只能取小于 k 这一侧的区域作为拒绝域，如图 6-2 所示，这就是“单边”的含义。

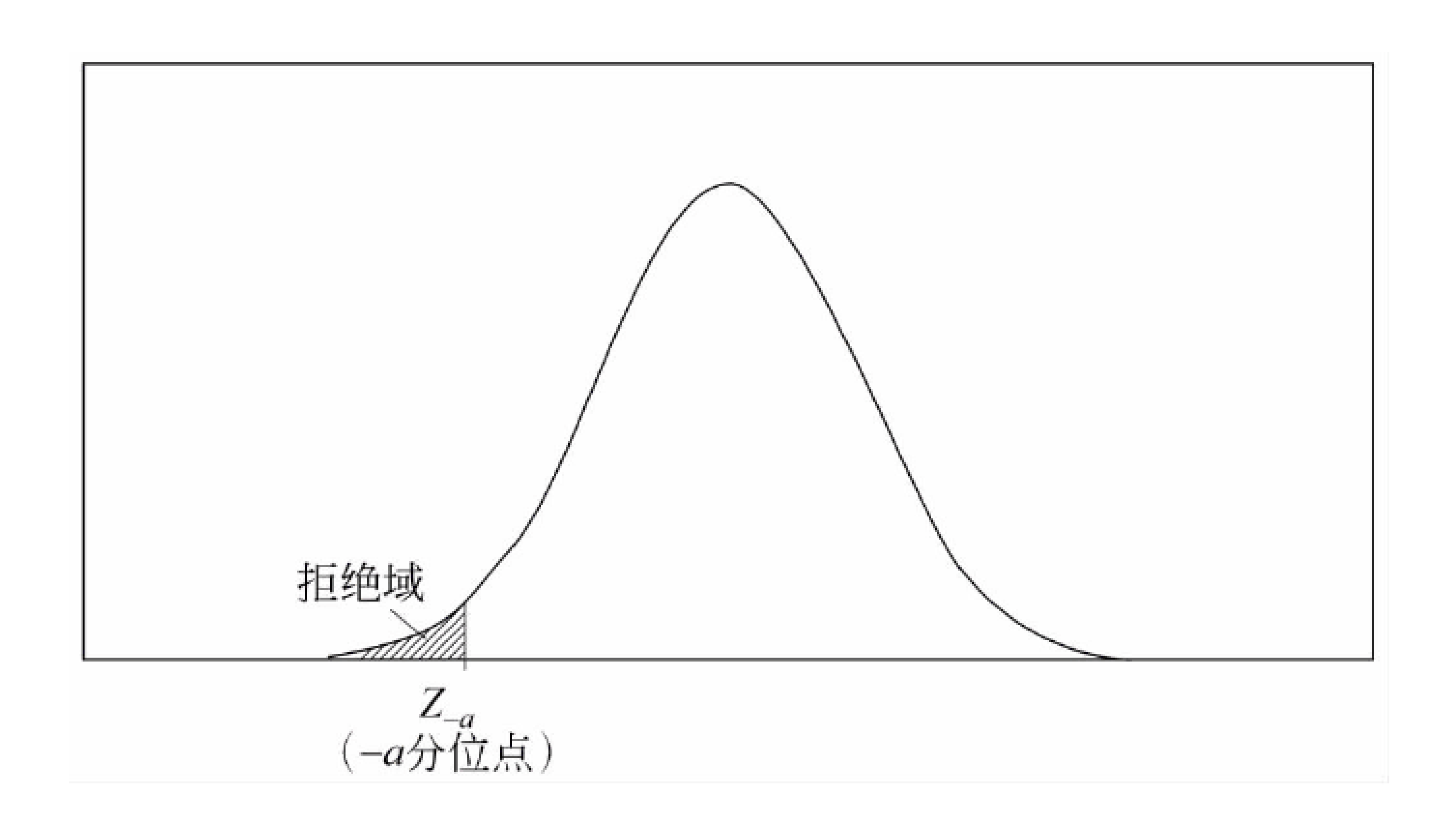


图 6-2　单边 Z 检验的临界点和拒绝域

$Z=(\overline{X}-0.1)/(0.9/\sqrt{30})$ 服从 $N(0,1)$ 标准正态分布，采样数据的均值为0.15，对应的 Z 值为 $(0.15-0.1)/(0.9/\sqrt{30})$，即 0.30。

取显著性水平 $\alpha=0.05$，此时，拒绝域是单边的，查表可得临界点为 $z_{-\alpha}=-1.65$，即拒绝域为 $Z<-1.65$。由于 0.30 没有落入拒绝域，我们接受 H_0 假设。

如果将两个对立假设改为：

H_0：$\mu\leqslant 0.1$

H_1：$\mu>0.1$

其余条件不变，则拒绝域的形式为

$$\overline{X} > k(k\text{ 为某个常数})$$

取显著性水平 $\alpha=0.05$，此时，拒绝域是单边的，查表可得临界点为 1.65，即拒绝域为 $Z>1.65$。由于 0.30 没有落入拒绝域，我们接受 H_0 假设。

上面的两个 H_0 假设包含自相矛盾的含义，我们却都接受了，看似存在矛盾，既然计算过程没有错误，那么问题一定出在前提条件上。每一次检验，我们都假定 X 服从 $N(\mu,0.9^2)$ 的正态分布，这个假设是进行 Z 检验的前提条

件，可是，0.9 只是我们从样本估计出的总体标准差，不一定是真正的总体标准差，如果我们事先不知道总体标准差，也可以进行假设检验——t 检验。

t 检验

已知主客场胜率差 X 服从正态分布 $N(\mu,\sigma^2)$，μ 和 σ 都是未知参数，此时，我们要验证 $\mu=\mu_0$ 是否正确，就需要使用 t 检验。

构造两个对立假设：

H_0：$\mu=\mu_0$

H_1：$\mu\neq\mu_0$

假设 H_0 成立，由于 σ 是未知变量，不能构造 Z 值进行检验，此时，我们利用“正态分布”一节中介绍过的 t 分布来构造一个变量 t。

已知 $X\sim N(\mu,\sigma^2)$，X_1，X_2，…，X_n 是 X 的 n 个样本，因此，$(n-1)S^2\sim X^2(n-1)$，其中 S^2 是 σ^2 的无偏估计。根据 t 分布的定义，可以构造变量 t，

$$t=(\overline{X}-\mu_0)/(S/\sqrt{n})$$

随机变量 t 服从 $t(n-1)$ 分布。与双边 Z 检验相似，在双边 t 检验中，只需要找到 t 分布的 $-\alpha/2$ 分位点和 $\alpha/2$ 分位点，就可以确定拒绝域，进而判断 t 值是否落入拒绝域。

我们使用 t 检验重新验证 μ 与 0.1 的关系。

构造两个对立的假设：

H_0：$\mu=0.1$

H_1：$\mu\neq0.1$

假设 H_0 成立，$t=(\overline{X}-0.1)/(0.9/\sqrt{30})$ 服从 $t(29)$ 分布，取 $\alpha=0.05$，可得 t 分布的 $-\alpha/2$ 分位点和 $\alpha/2$ 分位点分别为 -1.70 和 1.70，样本均值 0.15 对应的 t 值为 0.30，并未落入拒绝域中，因此我们接受 H_0 假设。

单边 t 检验与双边 t 检验类似，找到 $-\alpha$ 或 α 分位点即可确定拒绝域，这里不再赘述。

除了 μ，我们还可以针对 σ^2 进行假设检验，此外还可以对两个正态分布随机变量的期望之差 $\mu_1-\mu_2$ 进行假设检验，这些假设检验涉及 X^2 检验、F 检验等

更复杂的检验方法，但是基本思想和计算过程与 Z 检验和 t 检验类似，感兴趣的读者可以阅读盛骤、谢式千和潘承毅老师的《概率论与数理统计（第四版）》一书第八章。

最后，我要向读者做一个小小的“检讨”。在本节中，我们一直把主客场胜率差服从正态分布当作已知条件，这是值得质疑的。在写作本节前，我本应搜集大量样本或借用其他研究结果，对主客场胜率差服从正态分布做出验证，但是我没有做这个工作。虽然这并不会影响假设检验的学习，但是我依然要提醒读者，在实际应用假设检验时，随机变量是否服从正态分布需要谨慎判断。

6.3 反证法：无罪推定

假设检验背后隐含着的一个经典的证明方法——反证法。所谓反证法，是先假设求证的结论成立，再尝试从假设和已知条件中推理出相悖的结论，如果相悖的结论存在，说明假设是错误的，从而认定求证的结论不成立。

以双边 Z 检验为例，我们要求证明 $\mu=0.5$，首先假设 $\mu=0.5$ 成立，然后利用已知条件和采样数据进行推理，发现 $\mu=0.5$ 对应的 Z 值落入了拒绝域，这说明采样数据并不符合预期的正态分布，因此我们拒绝接受 $\mu=0.5$ 的假设，即 $\mu=0.5$ 不成立。

反证法是逻辑学中一个重要的证明方法，在很多领域都有应用，现代法律中的“无罪推定”原则是反证法的最佳例证。

无罪推定

法律的目标是维护正义，惩罚邪恶，每一个执法者都希望真正的罪犯得到法律的制裁。可是，在法庭断案时，执法者难免会犯两类错误：一类是错杀，即为无辜者定了罪；另一类是漏判，即真正的罪犯没有得到惩罚。这两类错误，哪一类更应该避免？恐怕大多数人会认为——都应该避免。可是说起来容易，做起来难，因为要避免这两类错误，需要遵从不同的判断逻辑——无罪推定和有罪推定。

无罪推定的原则是优先避免错杀，判断逻辑是，假设嫌疑人有罪，极力寻找推翻假设的证据，哪怕有一个证据能推翻假设，也不能判定嫌疑人有罪；有罪推定的原则是优先避免漏判，判断逻辑是，假设嫌疑人无罪，极力寻找推翻假设的证据，只要有一个证据能推翻假设，就可以判定嫌疑人有罪。

无罪推定虽然避免了错杀，却可能使凶手逃过法律的制裁，有罪推定虽然避免了凶手漏网，却可能使无辜的人蒙冤入狱。现代法律重视每个人的人权，因此选择了“宁可漏判，不可错杀”的无罪推定原则作为法庭判案的基本原则。

20 世纪末期发生的“辛普森杀妻案”是无罪推定的代表案件。1994 年 6 月 12 日深夜，洛杉矶西部一豪宅门前发现一男一女两具尸体。女性死者是著名黑人橄榄球运动员辛普森的前妻妮克・辛普森，男性死者是餐馆服务生郎・高曼，两人均遭利器割喉而死。案发后的凌晨，四名警察来到辛普森的住所，发现大量证据——门外的白色汽车染有血迹，车道上也有血迹，后院里有一只染有血迹的手套。辛普森在芝加哥酒店接到警方通知，清早赶回加州，几天之后，他被列为本案主要嫌疑犯，遭到逮捕。

庭审不久后开始。检方在开庭陈词中指控辛普森预谋杀妻，作案动机是嫉妒心和占有欲。离婚之后，辛普森对妮克与年轻英俊的男人约会非常吃醋，一直希望破镜重圆。案发当天，在女儿的舞蹈表演会上妮克对辛普森非常冷淡，使他萌动了杀机。服务生郎・高曼属于误闯现场，偶然被杀。法医鉴定表明，被害人死亡时间大约在 22：00～22：15。辛普森声称，当晚 21：40～22：50 他在家中独自睡觉，无法提供证人。辛普森豪宅中发现的沾有血迹的汽车和手套是重要证据。

看起来，一切的证据都表明辛普森是凶手，可是，随着庭审的进行，所谓的“证据”却遭到辩方律师的有力反驳。辩方认为，辛普森作为一个“业余杀手”，要实施谋杀理应用枪，割喉的实施难度很大而且容易留下大量血证，辛普森前妻妮克有吸毒史，此番很可能是遭遇黑手党的杀害，而割喉恰恰是黑手党杀手常用的杀人手法。案发现场的血证也存在诸多疑点，沾有血迹的袜子左右两侧的血迹竟然完全相同，这不可能是凶手穿着的袜子，更可能是被人涂抹上去的，辛普森豪宅后院的五滴被告血迹大小均匀、外形完整，也不合常理。更令人生疑的是，现场血迹中发现了浓度很高的螯合剂，案发之日，警方在辛普森的血样中添加了这种螯合剂，并曾携带血样返回案发现场。最后，辛普森被要

求当庭戴上沾有血迹的手套，可是辛普森折腾了很久也很难将手套戴上，辩方由此认为这只手套太小，根本不可能是辛普森的。

辛普森一案是当时美国社会白人与黑人对立的集中反映，辛普森虽然是黑人，却喜欢结交白人朋友，热衷于跻身富有的白人圈子，遭到大多美国公民的厌恶。庭审之初，陪审团成员普遍倾向于辛普森有罪，可是辩方对现场证据提出的质疑逻辑严谨、难以反驳，这些质疑渐渐动摇了陪审团的初始判断。1995 年 10 月 3 日，美国西部时间上午 10 点，辛普森案裁决即将宣布之时，整个美国一时陷入停顿，连同克林顿总统在内的 1 亿 4 千万美国人收看或收听了“世纪审判”的最后裁决——陪审团裁决结果：辛普森无罪。

虽然辛普森杀妻案已经过去了二十多年，但它依然是无罪推定的代表案件，我国法律中虽然没有陪审团制度，但也坚持无罪推定原则，下面，我们借用一部电影来亲身感受一下庭审上的无罪推定原则。

十二公民

国产电影《十二公民》翻拍自经典老片《十二怒汉》，讲述了十二个学生家长组成陪审团审理案件的全过程，是一部精彩的庭审题材电影，让我们跟随电影中的人物一起感受和学习庭审中的“无罪推定”原则。

“朝阳区某居民区内发生杀人案件，一名四十岁左右的河南籍男子被人刺死在家中，案发现场的场景被围观群众录像，视频一经上传，一小时内点击数已破十万，嫌疑人的姓名曝光后身份很快被网友人肉出来，此人现年 21 岁，是本市有名的富商之子，死者正是富二代的生父。不久前，检察院却做出了存疑不起诉的决定，再次将整个案件推到风口浪尖。富二代杀人案引起社会各界巨大反响，并引发了各大媒体甚至法学院的讨论热潮。”

影片围绕着一桩“富二代杀人案”展开。法学院以这桩知名的“富二代杀人案”为英美法律课的补考题目，邀请补考同学的家长和学校保安、小卖店店主等十二个“法律外行”组成模拟陪审团。在模拟庭审环节后，该陪审团需要在至少一个小时的时间里充分讨论，得到一致的结论——十二票全部赞成有罪或者十二票全部赞成无罪。

接下来，模拟陪审团的讨论正式开始。

团长发起第一轮投票，结果是十一票有罪，一票无罪。大家本以为这个讨论只是走走过场，赶紧投出个十二票有罪就结了，谁知 8 号陪审员偏偏投了无罪。而他的说法竟是："我是真觉得咱们应该讨论讨论。"其他几位陪审员顿时急了，轮番发言，试图说服 8 号陪审员，他们给出的理由是"网上铺天盖地的帖子都说人是富二代杀的""证据挺明显的""这就是个一清二楚的案子"。8 号陪审员的态度十分坚决："这事咱不想清楚，不说明白了，随随便便把手这么一举，就把这孩子往死道上这么一推，这，太快了。"——8 号陪审员表现出了一位陪审员应有的职业态度，对待有罪判罚要慎之又慎。此外，陪审员要坚持自己的独立判断，不能盲目接受媒体和网友的言论，在一些重大案件的审理过程中，陪审员甚至会"享受"与世隔绝的待遇，其目的就是让陪审员避免外界干扰，作出独立判断。

接下来，其他 11 位陪审员开始轮番表达意见，试图说服 8 号改判有罪。

10 号是个老北京人，他说："你得看这是什么人教育的孩子！这孩子的亲爹是河南一农民，蹲过大狱还离过婚，一个能把自个亲儿子给扔了的人他能是好人吗？这孩子的后爹也是河南一农民，也就十年的工夫，从负债累累到身价过十亿的药业大款，他这后爹要不干点违法乱纪的事儿，他能挣这么多钱吗?"——10 号的发言是典型的主观臆断，毫无事实根据，这是陪审员的大忌。

2 号是个胖墩墩的老好人，他笑着说："关于这个案子，我没什么说的。我就是觉得，这孩子有罪。因为从反证法的角度看，我们没法证明这个人不是他杀的啊。"——2 号很明显落入了有罪推定的逻辑，8 号马上纠正了他。

8 号说："我们根本不用证明不是他，只要证明证据中存有疑点。"——陪审团的职责是从证据中寻找疑点，试图排除嫌疑人的杀人嫌疑。

至此，十二位陪审员方才明确了陪审团奉行的无罪推定原则，在 8 号的指引下，陪审团开始整理证据。

本案的证据主要有如下 3 个。

证据 1：老头儿的证词

住在案发现场楼下的老头儿，在案发当天晚上 12 点 10 分的时候，听见楼上爷俩儿吵起来了，那个富二代大喊："我要杀了你!"一秒钟之后，他又听见，有人倒在地上了，老头儿赶忙起床跑到门口，15 秒左右，他在自家门口刚好看见那个富二代从楼梯上跑下来，走了。于是，老头儿赶忙打电话报警，警察来

了发现，死者的身上插了一把刀。

证据2：凶器

富二代一直在自己车上放着一把弹簧刀，这把刀与犯罪现场发现的凶器一模一样，案发后警察找不到富二代的刀，富二代说自己的刀丢了。富二代的这把刀被网友曝光在网上，它外观特别，还带有编号，看起来是一把私人定制的刀。

证据3：女人的证词

案发当晚，住在案发现场对面的女人躺在床上辗转反侧、无法入睡，她无意中透过驶过的城铁车窗，看见男孩捅了他的生父。

其他陪审员受到8号陪审员的感召，一同讨论起这三个证据，此前看似牢不可破的“铁证”暴露出了很多疑点。

疑点1：凶器

8号陪审员在网上花66元买了一把和凶器一模一样的弹簧刀，这说明这把刀并非私人定制，网友发布的照片很可能是PS处理过的。因此，案发现场的刀并不能与富二代构成必然联系。

疑点2：喊声

经陪审团估算，一列6节长的城铁驶过案发现场的窗口大约需要6秒钟，并且会发出巨大的轰隆声，案发现场离城铁很近，城铁通过时，住在楼下的老头儿理应听不清楼上发出的任何声音，自然也听不清富二代的喊声。

疑点3：时间

住在楼下的老头儿是瘸子，经陪审团现场模拟，老头儿从听见楼上有人倒地，到挪步至门口大约需要43秒，这与证词中的“十五秒左右跑到门口”相矛盾。

疑点4：刀口

富二代身高一米七二，死者身高一米八三，死者身上的刀是由上向下插入的，侧跳型弹簧刀一般的用法都是由下向上捅进去，更何况死者比富二代高了十一厘米，由上向下插入显得不合情理。

疑点5：目击

自称案件目击者的女人常常揉鼻梁，眼窝里有两个坑儿，喜欢眯眼看东西，这些细节说明这个女人很可能是近视眼，而试图睡觉的人是不会戴眼镜

的，一个没戴眼镜的近视眼能否看清几十米外的凶杀过程，令人怀疑。

至此，最后一个投有罪票的人也改判无罪，标志着陪审团最终达成了一致，从十一票有罪一票无罪，到十二票无罪，陪审团的每一个人都更加深刻地理解了庭审上的无罪推定原则。影片中 8 号陪审员的两句话最能说明无罪推定原则的内涵，他说：

"谁也不能随随便便宣布一个人有罪!"

"你手握生杀大权，杀了一个无辜的人，你跟凶手有什么区别?"

读者可以放下本书，看看这部电影，希望你也能从中领悟无罪推定的要义。

第 7 章

贝叶斯定理

导语：它曾遭受质疑，险些被遗忘；它的数学表达非常简单，却蕴藏着深刻的概率思想；它在医学、刑侦、博彩等领域得到广泛应用，连机器学习算法中也有它的一席之地。它就是概率统计中最具实践意义的贝叶斯定理。

7.1 牧师贝叶斯：深藏功与名

在概率统计中，大数定理是最具理论意义的定理，贝叶斯定理则是最具实践意义的定理。贝叶斯定理不仅在医学、刑侦、博彩等领域得到广泛应用，还衍生出了朴素贝叶斯分类器、贝叶斯网络等新方法，在机器学习、不确定性推理等领域也占有重要位置。下面我们就来认识一下贝叶斯定理的创始人——“牧师”贝叶斯。

牧师贝叶斯

托马斯·贝叶斯(约1701—1761)是一位受人尊敬的英格兰长老会牧师，

同时也是英国皇家学会会员。他相信神是完美的,这世界上之所以还有邪恶和苦难,是因为人类对自然和宇宙的了解还不够,所以我们要不断探索宇宙的规律。业余时间里,他喜欢研究一些逻辑和概率方面的问题。当时,人们对概率的认识还十分肤浅,如何理解“逆概率”尚无定论,这引起了贝叶斯的兴趣。

常见的概率问题往往是这样的:已知袋子里有 5 个红球、8 个蓝球,闭上眼睛拿出一个,拿到红球的概率是多少?这是“正概率”问题。“逆概率”问题与之相反:袋子里有很多红球和蓝球,从中随意拿出 5 个,发现 3 个是蓝球、2 个是红球,那么袋子里红球和蓝球的比例可能是怎样的?

贝叶斯利用业余时间对“逆概率”问题做了很多研究,并撰文记录下了自己的研究成果。可惜贝叶斯提出的理论与当时的主流统计观点相左,他的研究成果因此遭到了冷落。贝叶斯死后两年,他的好友理查德·普莱斯将他的文章寄给了英国皇家学会,这篇贝叶斯定理的开山之作方才公之于众。

贝叶斯撰写的文章是《机会问题的解法》(*An essay towards solving a problem in the doctrine of chances*),文章的表达清晰明确,将“逆概率”问题以点、线、面的方式直观的呈现出来,并在解答过程中提出了贝叶斯公式。更让人钦佩的是,文章中有关概率的表述十分准确,却没有使用任何概率相关的数学表达式,对一个“业余”的数学爱好者来说实属不易。

后来,法国数学家拉普拉斯把贝叶斯定理总结为一个简洁的数学表达式,从此贝叶斯定理被人们接受,并得到了越发广泛的应用。

贝叶斯定理

贝叶斯定理之所以得到广泛应用,与其简洁的表达式不无关系。在“条件概率”一节我们学习过如下公式:

$$P(AB) = P(A \mid B) \cdot P(B)$$

$$P(B) = P(AB) + P(\bar{A}B)$$

式中,A 和 B 分别表示两个随机事件,$\bar{A}$ 表示 A 的逆事件,即事件 A 不发生。

将这个公式做简单的数学变换,便可以得到贝叶斯定理的表达式:

$$P(A \mid B) = P(B \mid A) \cdot P(A)/\ P(B \mid A) \cdot P(A) + P(B \mid \bar{A}) \cdot P(\bar{A})$$

这个公式看起来并不“简洁”,这是概率的表达符号带给你的错觉,我们用

x 表示 $P(A)$，用 y 表示 $P(B|A)$，用 z 表示 $P(B|\bar{A})$，便可以把贝叶斯定理改写成下面的形式：

$$P(A \mid B) = xy/[xy + z(1-x)]$$

改写过后，贝叶斯定理显得简洁多了，它的含义也变得清晰了。要计算条件概率 $P(A|B)$，只需要知道 $P(A)$、$P(B|A)$和 $P(B|\bar{A})$。明明是把计算量变大了，为什么说“只需要”？因为计算难度降低了。在很多现实问题中，$P(A|B)$往往难以直接计算，而 $P(A)$、$P(B|A)$和 $P(B|\bar{A})$却可以计算（或估算）出来，贝叶斯定理的奥秘就在于此。

我们一起来看下面这个例子——“你身上有她的香水味”。

你和丈夫新婚刚刚半年，正是如胶似漆的时候，丈夫却忽然因公出差一个月。你盼着，盼着，一个月后，终于把丈夫盼回来了。可是，就在你拥抱归来的丈夫时，你的鼻子却嗅到了不该嗅到的气味——女人的香水味。你知道，女人的鼻子永远不会犯错，这一定是另一个女人留下的味道！你无法排解心中的难过和纠结：难道丈夫出轨了？

下面我们用贝叶斯定理计算“丈夫出轨的概率”。

设随机事件 A 表示丈夫出轨，随机事件 B 表示丈夫身上有其他女人的香水味，我们的计算目标是 $P(A|B)$。根据贝叶斯定理，我们要分别计算 $P(A)$、$P(B|A)$和 $P(B|\bar{A})$三个概率值。$P(A)$表示在没有任何已知条件时丈夫出轨的概率，假设你相信自己的丈夫很专一，$P(A)=1\%$，这个概率相当低。$P(B|A)$表示，在丈夫出轨的前提下香水味出现的概率，这个概率一定很高，但是你丈夫并不傻，出轨之后一定会试图洗白自己，综合来看，这个概率可以设为 60%。$P(B|\bar{A})$表示丈夫没出轨的前提下香水味出现的概率，也许是结伴女同事在丈夫身上留下的，可是丈夫所在的公司女同事很少，这种情况出现的概率很低，估计只有 10%。

估算出了 $P(A)$、$P(B|A)$和 $P(B|\bar{A})$三个概率值，便可以代入贝叶斯公式中，得到丈夫出轨的条件概率为 $P(A|B)=6\%$，你长舒一口气，丈夫出轨的概率还是很低的。

在这里例子中，你对丈夫本人的极度信任十分关键，如果你对他的信心稍有动摇，比如 $P(A)=10\%$，其他条件都不变，丈夫出轨的条件会暴涨到 40%！

$P(A)$被称为先验概率，在很多实际问题中，$P(A)$只能借助主观推测，这

也是贝叶斯定理自提出之日起就为人质疑的一点。为了摒除主观推测的干扰，统计学家们提出了“频率主义”。

频率主义 vs 贝叶斯定理

频率主义是统计学中的一种思想，它力图摒弃任何主观推测，站在绝对客观的角度搜集数据，并用严谨的数学模型概括数据的特征。推崇频率主义的统计学家会无止境地搜集数据，对数据的概率分布做出统计假设，再用假设检验来验证统计假设是否正确。然而，这种思想有明显的“完美主义”倾向，容易脱离现实。比如，频率主义始终在力图解释抽样误差，它希望找到、也相信能找到一个通用的方法来计算抽样误差，从而消除统计偏差，却始终未能实现。在当下的大数据时代，样本近乎就是总体，抽样误差不再存在，然而即便使用总体中的全部数据，也往往无法做出合理的统计推断，因为数据越多，噪声也越多。

“不识庐山真面目，只缘身在此山中”，我们生活在世界上，很难找到绝对客观的视角来看待世界。贝叶斯定理的确包含主观推断，可是定理中的主观推断通常是以经验数据作为参考，即便主观推断最初可能出错，还可以借助经验数据的搜集，不断迭代和更新主观推断，一步步接近真相。相比于频率主义，贝叶斯定理更加接地气，在医学、博彩、刑侦等很多实际问题中，贝叶斯定理都发挥了不可替代的作用。

7.2 赌神贝叶斯：一赌定终身

每年 11 月，NBA 会迎来新赛季，身住洛杉矶的富豪哈若拉波斯 · 乌尔加利斯也开始了新赛季的工作。每天晚上，他会对着家中的 5 台高清平板电视机，同时观看 5 场比赛，更新自己的比赛数据库，并为下一次下注做好准备。乌尔加利斯是一位“NBA 职业赌客”，他不仅没有因为赌博输钱，正相反，他因为赌博而成了千万富翁。年景不好时，乌尔加利斯也能赚到 100 万美元，年景好时，赚上三四百万美元也不在话下。乌尔加利斯是怎样成为“赌神”的？一

次疯狂的下注成就了他。

乌尔加利斯出生于加拿大，他的父亲曾经坐拥 300 万美元身家，却因嗜赌而破产。乌尔加利斯继承了父亲的基因，从小就对赚钱拥有强烈的欲望，他在大学期间做了很多份兼职，赚到了人生的第一笔 8 万美元。就在大四这一年，他意外地陷入一场赌局中而无法自拔，这场赌局就是——湖人队能否夺得 1999－2000 赛季 NBA 总冠军。

那一年，湖人队聘用了"禅师"菲尔·杰克逊当教练，阵中坐拥全联盟最强中锋"大鲨鱼"奥尼尔，天赋异禀的明星后卫科比，以及一批实力不俗的角色球员。但这似乎并没有让拉斯维加斯的赌客们高看湖人队。上个赛季，湖人队一直风波不断，年轻气盛的科比与奥尼尔爆发了矛盾，球队三次换帅，最终在季后赛被马刺队横扫出局。本赛季常规赛第三场，湖人队输给了经验老到的开拓者队，更糟糕的是，奥尼尔在场上发飙被裁判驱逐出场，一切似乎都是上个赛季的重演，湖人队要在残酷的季后赛中战胜马刺队和开拓者队简直是天方夜谭。

1 赔 7.5——拉斯维加斯的庄家调高了湖人队夺冠的赔率，湖人队的家乡媒体《洛杉矶时报》也看衰湖人队的前景。乌尔加利斯是个喜欢挑战权威的小伙子，他很欣赏菲尔·杰克逊的执教风格，并且深信湖人队不会如此不堪，他决定赌一把！除了必需的生活费外，他把自己打工赚来的 8 万美元全部下注湖人队夺冠。以这个赔率计算，如果湖人队最终问鼎冠军，乌尔加利斯会赚到 50 万美元！

他真的赚到了！

那一年，湖人队在常规赛取得了 61 胜 21 负的不俗战绩，但是到了季后赛，他们遭遇了国王队和开拓者队的强力阻击。对阵开拓者队的西部决赛被拖入一场决胜的抢七大战，第三节临近打完时，湖人队居然在自己的主场落后 16 分之多，参照历史数据，湖人队翻盘的概率不足 20%。即便难以取胜，湖人队也决不能在家乡父老面前缴械投降。最后一节，在主场球迷的疯狂呐喊中，湖人队众志成城，将比分一步步迫近，并借助科比的两粒罚球一举反超，最终乘势拿下了比赛胜利。神奇的主场优势助湖人队完成了不可思议的大逆转！总决赛中，湖人队轻松战胜步行者队，顺利夺冠！

这一赌开启了乌尔加利斯的赌神之路！乌尔加利斯有了足够的资金，能

够承担起小额投注的风险，于是他开始尝试成为NBA职业赌客。他会在某个NBA比赛日同时下注三、四场比赛，并不断矫正自己的投注策略。后来，他开始搜集各场比赛的相关信息，寻找比赛过程、比赛结果与赛前各种信息之间的关联，比如，某个球员的绯闻女友在推特上暗示今晚要和这个球员去夜总会，可这个球员的心思根本不在比赛上，他在当晚比赛中的表现八成会很糟糕。乌尔加利斯不断优化自己的投注策略，搜集的信息也越来越多，现在他经营着一家球探机构，雇用球探来搜集比赛信息，并详细分析各支球队的攻防战术。此外，他还关注超过100位NBA球员和教练的推特账号，从中搜集球员在赛前赛后的各种言论，球队教练在赛前新闻发布会上的措辞也是乌尔加利斯重点关注的内容。这些信息最终会输入乌尔加利斯的一个计算机仿真程序中，帮助他模拟比赛结果。

乌尔加利斯认为，没有什么理论能够准确地预测未来，但是，未来却是由当下不断发生的各种事件共同决定的，比赛的胜负如同股票指数一样变幻莫测，每一个利好或利空消息都会产生或大或小的影响，成功的赌徒从纷繁复杂的消息中去除噪声，判断这些消息产生的影响到底有多少。我们很难把乌尔加利斯的方法抽象成某种数学理论——当然，他也不会告诉我们他的方法是什么——但我们知道，隐藏在他的方法背后的思想正是贝叶斯定理。

湖人队的夺冠概率

时间回溯到1999年11月，乌尔加利斯加入了“湖人队能否在2000年夺冠”的赌局中，当时NBA常规赛刚刚进行了12场比赛，湖人队的战绩是8胜4负，根据这个已知条件，我们来计算湖人队夺冠的概率有多少。

设定随机事件 A 表示湖人队夺冠，随机事件 B 表示前12场比赛8胜4负，求解 $P(A|B)$ 的过程如下所述。

(1) 估算先验概率 $P(A)$。联盟至少有2支球队的实力与湖人队相当，而且湖人队在上个赛季的季后赛被马刺队横扫出局，因此湖人队本赛季夺冠的概率并不会高，可以设为20%。

(2) 估算 $P(B|A)$ 和 $P(B|\overline{A})$。我们可以查阅 *NBA* 的历史资料，以最近几个赛季的联盟强队的战绩为参考，估算湖人队夺冠和不夺冠的情况下，出现

8胜4负战绩的概率，假设它们分别为60%和50%。

(3) 根据贝叶斯定理，可以计算得到湖人队夺冠的概率为：

$$P(A \mid B) = P(B \mid A) \cdot P(A) / \ P(B \mid A) \cdot P(A) + P(B \mid \overline{A}) \cdot P(\overline{A})$$
$$= \cdots = 23\%$$

可见，湖人队的开局战绩并没有拉低夺冠概率。按照1赔7.5的赔率，在下注8万美元赌湖人队夺冠的情况下，如果湖人队夺冠可以得到52万美元的净利润，如果湖人队没能夺冠，将失去8万美元，因此，乌尔加利斯在这个赌局中的获利期望是：

$$E(\text{获利}) = 23\% \times 52 + 77\% \times (-8) = 5.8(\text{万美元})$$

可见，对乌尔加利斯来说，这个赌局是有利可图的。虽然他最终凭借这个赌局一举成为富翁，但他清楚，概率只有在多次试验中才会发挥效应，要坚持使用贝叶斯定理的思想，同时不断优化自己的投注策略，才能成为一个成功的赌客。虽然乌尔加利斯靠赌博赚到了很多钱，但是他的下注正确率也只能达到57%，去掉庄家的“抽头”，只剩下非常小的获利空间，正如大数定理对庄家的作用一样，只要能一直保持57%的正确率，看似微小的盈利就会像滚雪球一样积少成多，这就是成功赌客的秘诀。

7.3　死神贝叶斯：连环恐怖袭击

“战争”是人类文明史上的高频词，距今仅100年前，第一次世界大战正在无情地吞噬着欧洲大陆上的生命，1918年11月，第一次世界大战结束，留下了1 000万人丧生、2 000万人受伤的惨烈数字。20年后，法西斯头目阿道夫·希特勒和墨索里尼联手挑起了第二次世界大战，经过6年的苦战，世界反法西斯同盟战胜了法西斯轴心国，却付出了7 000万人死亡、上亿人受伤的惨痛代价。经过这次波及全球的世界大战，反法西斯同盟的各国绝不希望再次爆发大规模战争，由此成立了联合国安理会，共同维护世界和平和安全。此后的数十年，地球告别了大规模战争，全世界的经济、科技日益发展繁荣，然而，在和平发展的表象下，一股黑暗势力正在滋生和扩张，它就是——恐怖主义。

恐怖主义，是指恐怖组织对非武装人员(通常是平民)进行暴力袭击，以实现其政治或宗教目的，袭击方式包括制造爆炸、劫机、绑架、暗杀等。恐怖组织包括极左(右)翼恐怖主义团体、极端的宗教主义或种族主义组织等，他们往往有非常明确的政治或宗教诉求，拥有强大的资金支持，能够从世界各地招募成员，并将他们训练成无比虔诚的教徒，然后针对特定目标发动自杀式恐怖袭击。当下规模最大、最活跃的恐怖组织非 IS 莫属。

伊斯兰国(Islamic State，IS)，前称伊拉克和大叙利亚伊斯兰国(Islamic State of Iraq and al Shams，ISIS)，是一个活跃在伊拉克和叙利亚的极端恐怖组织。在 2003 年，伊拉克战争期间，IS 还只是“基地”组织的一个分支，自 2011 年，“基地”组织头目本·拉登被美军击毙，美军随即撤出伊拉克，此后 IS 组织迅速壮大，趁叙利亚内战之机进驻叙利亚，于 2014 年 2 月宣布“建国”，“定都”叙利亚城市拉卡。随后，IS 脱离“基地”组织独立发展。2014 年 6 月，IS 头目巴格达迪宣布在伊拉克和叙利亚建立伊斯兰帝国，不久之后，阿富汗恐怖组织塔利班宣布效忠 IS，助其建立全球性的伊斯兰帝国。

2014 年 9 月，美国联合英国、法国等 54 个国家和欧盟、北约等国际联盟发动了对 IS 的军事打击，接连遭遇空袭和地面袭击的 IS 组织开始疯狂回击，他们先后对法国、英国和日本公民实施斩首，并将斩首视频公之于众。此后，他们对欧洲实施了两起骇人听闻的恐怖袭击。第一起是法国巴黎的“11·13”恐怖袭击，2015 年11 月 13 日晚，法国首都巴黎的市中心连续发生多起枪击和爆炸事件，造成 128 人死亡，99 人重伤，IS 随后宣布对该事件负责，并称其为一个“奇迹”；第二起是比利时布鲁塞尔的“3·22”恐怖袭击，袭击发生前四天，布鲁塞尔警方刚刚逮捕了巴黎恐怖袭击案的一名在逃嫌犯萨拉赫·阿卜杜勒·萨拉姆，引发了布鲁塞尔的穆斯林居民集体骚乱，四天后的 2016 年 3 月 22 日晚上，布鲁塞尔机场发生枪击和爆炸，随后位于欧盟委员会附近的地铁站发生爆炸，该事件造成至少 34 人死亡，IS 随后宣布对该事件负责。

从 20 世纪末到现在，恐怖组织制造的恐怖袭击数不胜数，如表 7-1 所示。这些恐怖袭击大多为连环袭击，这是巧合还是必然？下面我们用贝叶斯定理来揭秘连环恐怖袭击的秘密。

表 7-1　全世界近 30 年的恐怖袭击案件

发生时间	恐怖袭击事件	伤亡情况
1988 年 12 月 21 日	洛克比空难：一枚炸弹在美航 103 班机上被引爆	机上 259 人和地面 11 人死亡
1995 年 3 月 20 日	在日本东京交通最繁忙的 3 条地铁的 15 个车站同时发生毒气事件	10 人死亡，75 人重伤
1998 年 8 月 7 日	“基地”组织用炸弹袭击了美国在肯尼亚首都内罗毕和坦桑尼亚港口城市达累斯萨拉姆的大使馆	224 人死亡
2001 年 9 月 11 日	“9・11”事件：自杀式炸弹袭击者劫持民航客机撞向世贸中心和五角大楼	2 977 人死亡
2002 年 10 月 12 日	印度尼西亚度假胜地巴厘岛上连续发生两起炸弹爆炸	202 人死亡
2003 年 8 月 29 日	伊拉克南部清真寺发生汽车炸弹爆炸	100 多人死亡，200 多人受伤
2004 年 3 月 11 日	马德里爆炸案：西班牙首都马德里发生 3 列旅客列车连环爆炸事件	至少 198 人死亡，约 1 800 人受伤
2005 年 7 月 7 日	伦敦市中心金融区的地铁站和两辆巴士相继发生爆炸	52 人死亡，700 多人受伤
2006 年 7 月 11 日	在印度孟买下班繁忙时间发生的 7 次连环爆炸炸毁多个火车车厢	187 人死亡
2008 年 8 月 4 日	两名暴恐分子在中国新疆维吾尔自治区喀什市驾车袭击边防官兵并引爆爆炸物	17 名官兵殉职，15 人受伤
2014 年 5 月 6 日	“博科圣地”组织对尼日利亚东北部一座边境城市发动袭击	约 300 人死亡
2015 年 11 月 13 日晚	法国巴黎发生连续枪击和爆炸	至少 197 人死亡
2016 年 3 月 22 日晚	比利时布鲁塞尔扎芬特姆国际机场出发大厅发生爆炸，随后欧盟总部附近地铁站发生爆炸	至少 34 人遇难

连环袭击不是巧合

对待小概率事件，统计数字不仅无用，而且会使人麻木，震惊全球的“9・11”事件是最好的例证。2001 年 9 月 11 日，“基地”组织的恐怖分子劫持了四架大型客机，其中两架撞击了世贸中心的南楼和北楼，一架撞击了美国国防部五角大楼，还有一架最终坠毁。“9・11”事件彻底挑战了美国安保部门的“想象力”。在“9・11”事件发生前，几乎没人会想到，恐怖分子会驾驶飞机撞

击世贸中心大楼。因为统计数字告诉我们，在“9·11”事件前的25 000天里，曼哈顿上空一直有飞机通航，但是只发生过两次类似“幢楼”的事件，因此，从时间来衡量，“飞机撞大楼”发生的概率只有0.008%，如果按照飞机架次来衡量则更低。与此同时，另一组数据却没有受到应有的重视，自1995年起，全球的自杀式袭击数量大幅增加，2000年迎来了最高峰——39起，而且早在1998年，“基地”组织就曾企图用飞机撞击世贸中心大楼，但未能得逞。北美航空航天防御司令部曾提议进行一次有关“被劫客机袭击五角大楼”的军事演习，却因这一想法太不现实而未被采纳。整个美国都被现实束缚了想象力。

从历史统计数据上看，“恐怖分子驾机撞世贸中心大楼”的确是小概率事件，但是在特定的时期、特定的条件下，这一事件却未必是“小概率”的，尤其是在类似事件已经发生的情况下。当恐怖分子驾驶第一架被劫客机撞上世贸中心大楼的时候，这件事瞬间便不再是小概率事件了，不仅如此，“飞机再次撞击世贸中心大楼”几乎是一定的！一切都归因于贝叶斯定理。

设随机事件A表示“恐怖分子驾机撞世贸中心大楼”，随机事件B表示“飞机第一次撞击世贸中心大楼”，$P(A)$、$P(B|A)$和$P(B|\bar{A})$如表7-2所示。由于美国历史上从未发生过类似事件，所以我们把先验概率$P(A)$设成0.005%，根据贝叶斯定理，$P(A|B)=38\%$。也就是说，仅仅“恐怖分子驾机撞击世贸中心大楼”这一个事件的发生，就让“飞机撞击世贸中心大楼”这一事件的先验概率从0.005%暴涨到38%！

表7-2　已知恐怖分子驾机撞击世贸中心大楼时的贝叶斯定理

事　　件	计算公式	概率(%)
恐怖分子驾机撞世贸中心大楼	$P(A)$	0.005
已知恐怖分子驾机撞世贸中心大楼，飞机第一次撞击世贸中心大楼	$P(B\|A)$	100
恐怖分子未驾机撞击世贸中心大楼的情况下，飞机第一次撞击世贸中心大楼(意外事故)	$P(B\|\bar{A})$	0.008
已知飞机第一次撞击世贸中心大楼的情况下，恐怖分子驾机撞世贸中心大楼	$P(A\|B)$	38.46

更要命的还在后头！

设随机事件A表示“恐怖分子再次驾机撞世贸中心大楼”，随机事件B表示“第二架飞机撞上世贸中心大楼”，$P(A)$、$P(B|A)$和$P(B|\bar{A})$如表7-3所

示,在先验概率 38%的情况下,$P(A|B)$居然高达 99.99%！这似乎应了中国的那句老话——祸不单行!

表 7-3　已知恐怖分子第二次驾机撞击世贸中心大楼时的贝叶斯定理

事　　件	计算公式	概率(%)
恐怖分子再次驾机撞世贸中心大楼	$P(A)$	38.46
已知恐怖分子再次驾机撞世贸中心大楼,第二架飞机撞上世贸中心大楼	$P(B|A)$	100
恐怖分子未再次驾机撞击世贸中心大楼的情况下,第二架飞机撞上世贸中心大楼(意外事故)	$P(B|\bar{A})$	0.008
已知第二架飞机撞上世贸中心大楼的情况下,恐怖分子驾机撞上世贸中心大楼	$P(A|B)$	99.99

贝叶斯定理为我们开启了另一个视角,去看待地震、瘟疫、金融危机等"小概率事件"。统计数字只能告诉我们,这些事件极少发生,可是这没有实践意义。事实是,当某些相关事件发生时,小概率事件很可能会变成普通事件,甚至必然事件！所以,在对待小概率事件时,最具实践意义的做法是,不断搜集相关信息,不断更新事件发生的概率,只有这样才能做到有备无患。

7.4　神探贝叶斯：嫌疑人 X的献身

1841 年,美国作家爱伦·坡发表了小说《莫格街凶杀案》,这部小说被世界公认为侦探小说的开山之作,从那以后,侦探小说常常成为欧美畅销书的代名词。在英国,阿瑟·柯南道尔创作的《福尔摩斯探案集》可谓无人不知,阿加莎·克里斯蒂创作的《无人生还》《尼罗河上的惨案》等作品也是家喻户晓,在美国,埃勒里·奎因的《希腊棺材之谜》《X 的悲剧》等作品都是经典之作。除了欧美,侦探小说在另一个国家也渐渐崭露头角,这个国家就是我们的近邻——日本。

1923 年,作家江户川乱步发表处女作《两分钱硬币》,从此拉开了日本侦探小说的序幕,江户川乱步也被奉为日本侦探小说的鼻祖。在日本,侦探小说被称为推理小说,这是因为日本早期的侦探作品非常看重严密的推理过程,这类作品也被归为"本格派",江户川乱步正是"本格派"的杰出代表。第二次世界

大战后，日本走上艰难的重建之路，推理小说也随之迎来了一次转型，“社会派”推理小说出现了，代表作家是松本清张。“社会派”推理小说不再局限于案情推演，它探究犯罪的社会原因，以此揭示社会的阴暗面，折射出人们内心潜在的苦闷和矛盾。“社会派”推理小说在日本经久不衰，最终迎来了它的集大成者——东野圭吾。相信读者们都看过或听过东野圭吾最著名的作品《白夜行》，遗憾的是，《白夜行》在当年并未获得日本推理界的最高奖项“直木奖”，东野圭吾并未放弃，笔耕不辍，终于在 2006 年摘得梦寐以求的“直木奖”，助他获得该奖的是他的另一部代表作《嫌疑人 X 的献身》。

接下来，我们就跟随着《嫌疑人 X 的献身》的情节，看看推理小说中的神探是如何应用贝叶斯定理的思想来推演案情的。

案情推演

《嫌疑人 X 的献身》是一个多人物视角的推理小说，我们对原著做一次改写，以侦探汤川学的视角复述这个故事。

汤川学，后文简称汤川，是一名大学老师，因为他善于逻辑推理，因此草薙警官每每遇到难办的案子便向他请教。三月十一日，帝都大学物理学科第十三号研究室内，汤川正在和草薙下西洋棋。草薙忽然接到电话，有突发案件，于是前去调查。几天后，草薙带着疑惑再次拜访汤川，并向汤川交代了案情。

三月十一日上午，一位老人在旧江户川的堤防边跑步，看到地上塑胶布的一端露出看似人脚的东西，他遂战战兢兢地掀起塑胶布，竟发现了一具尸体！警方的现场取证结果如下：尸身全裸，惨遭毁容，手指被烧过，指纹遭到破坏。死者为男性，脖子上有勒痕，此外没有明显外伤。尸体旁边扔下了一辆崭新的脚踏车，两个轮胎都被人放了气，车上有登记编号，车把上留有指纹。在距离尸体大约一百公尺处，发现了疑似被害者的衣物，衣物塞在一斗深的桶子中，部分遭到焚烧，包括外套、毛衣、长裤、袜子和内衣。

警方对现场证据做了深入调查，有如下发现。

（1）死者是被人往上拉扯勒死的，凶器很可能是电线，比如电热器常用的那种空心麻花绳式的电线。

（2）死亡时间是三月十日晚六点到十点。

(3) 尸体旁的脚踏车是三月十日晚在堤防附近的车站被偷的,脚踏车上留有死者的指纹。

(4) DNA 鉴定结果显示,部分烧毁的衣物的确是死者的。

为了确认死者身份,警方在全城搜集失踪者信息,最终在一家旅店的客房里发现了死者的毛发和指纹,店主确认该住户在三月十日晚上之后再没回旅店。警方因此确认死者名叫富坚慎二。警方对富坚慎二展开调查,他们发现,富坚曾是销售进口车的业务员,后因挪用公款被公司开除,后来富坚和妻子离了婚,但仍一直对前妻纠缠不放。警察随即登门拜访了富坚的前妻花岗靖子。花岗靖子有一双“大大的黑眼珠”,“是个脸蛋小巧的女人”,她和富坚离婚已五年,很少来往。据花岗靖子说,三月十日晚上,她和女儿六点半出门去看电影,然后在同一栋大楼里的拉面店用餐,接着又去 KTV 唱歌,十一点之后才到家。

案情调查至此,凶杀案的唯一嫌疑人是花岗靖子,下面我们用贝叶斯定理来分析花岗靖子作案的可能性。

我们先来分析案发地点。尸体在堤防边发现,并不代表案发地点就是堤防边,因此,存在两种可能结果。

A_1:案发地点是堤防边;

A_2:凶手杀人后运尸至堤防边。

在未引入其他证据前,我们可以假定这两种情况出现的概率各为 50%。

然后我们引入两个证据。

B_1:脚踏车上留下了死者的指纹;

B_2:死者的手被砸烂,指纹被破坏。

如果案发地点就在堤防边,那么死者应当是骑着脚踏车来到堤防边,而后被害,凶手砸烂死者的手是为了破坏指纹,而后可能因为紧张忘记了脚踏车上留有死者指纹,因此 $P(B_1B_2|A_1)$ 大约为 80%;如果凶手杀人后运尸至堤防边,那么凶手在脚踏车上留下死者的指纹,却又为了销毁死者的指纹砸烂死者的手,这是明显矛盾的两个行为,除非凶手逻辑混乱或精神失常,因此 $P(B_1B_2|A_2)$ 大约为 5%。

我们知道,$\overline{A_1}$ 就是 A_2,$\overline{A_2}$ 就是 A_1,因此,利用贝叶斯定理便可以计算得到:

$$P(A_1 \mid B_1B_2) = P(B_1B_2 \mid A_1) \cdot P(A_1) / \ P(B_1B_2 \mid A_1) \cdot P(A_1) +$$

$$
\begin{aligned}
& P(B_1B_2 \mid \overline{A_1}) \cdot P(\overline{A_1}) \\
& = P(B_1B_2 \mid A_1) \cdot P(A_1) / \ P(B_1B_2 \mid A_1) \cdot P(A_1) + \\
& \quad P(B_1B_2 \mid A_2) \cdot P(A_2) \\
& = 94.1\%
\end{aligned}
$$

$$
\begin{aligned}
P(A_2 \mid B_1B_2) & = P(B_1B_2 \mid A_2) \cdot P(A_2) / \ P(B_1B_2 \mid A_2) \cdot P(A_2) + \\
& \quad P(B_1B_2 \mid \overline{A_2}) \cdot P(\overline{A_2}) \\
& = P(B_1B_2 \mid A_2) \cdot P(A_2) / \ P(B_1B_2 \mid A_2) \cdot P(A_2) + \\
& \quad P(B_1B_2 \mid A_1) \cdot P(A_1) \\
& = 5.9\%
\end{aligned}
$$

因此，我们几乎可以断定案发地点就是堤防边。接下来，我们以案发地点在堤防边为前提条件，计算花岗靖子是凶手的概率。由于花岗靖子是唯一的犯罪嫌疑人，所以假定花岗靖子是凶手的概率为80%，即先验概率 $P(A)=0.8$。与凶手关系最大的线索是作案手法，死者脖子上的勒痕显示，死者是被人用电线之类的东西往上拉扯勒死的，我们把这条线索记为事件 B。花岗靖子身高160厘米，是个身材纤细的弱女子，死者身高170厘米，并非孱弱之人，而且死者并未服用任何麻醉类药物，因此，假如花岗靖子是凶手，这样的作案手法实在难以理解，$P(B|A)$ 大约为5%。我们假定 $P(B|\overline{A})$ 为50%，可以计算得到

$$
\begin{aligned}
P(A \mid B) & = P(B \mid A) \cdot P(A) / \ P(B \mid A) \cdot P(A) + P(B \mid \overline{A}) \cdot P(\overline{A}) \\
& = 28.6\%
\end{aligned}
$$

仅是这一个线索便将花岗靖子是凶手的概率降低到28.6%。警察还验证了花岗靖子母女的电影票票根，上面的确留有二人的指纹，KTV的服务生也在当天晚上见到了母女二人，这一切都使花岗靖子是凶手的概率不断降低。因此，警方怀疑花岗靖子有男性共犯，案发过程可能是，花岗靖子将富坚引到堤防边，然后由男性共犯将其杀害，至于未完全烧毁的衣物和脚踏车，可能是二人急于逃跑导致的。

在交代完案情后，草薙警官提起了一个人——达摩石神。石神住在花岗靖子隔壁，草薙警官走访花岗靖子时刚好碰到石神，便向他了解花岗靖子的情况。汤川学听到石神的名字，不禁回忆起了大学时的往事。汤川和石神是京

都大学的校友，汤川主攻物理学，石神主修数学，他们俩都是“学霸”，不同的是，汤川更加博学，石神则沉迷于数学世界。汤川对石神在数学方面的造诣十分欣赏，虽然与石神的交流不算多，但他依然能感受到二人之间是彼此理解的。毕业后，汤川选择了留校，与石神一别就是二十多年。此番听到这位“知音”的消息，有些欣喜，当晚便登门拜访，没想到汤川却在会面期间觉察到了石神对靖子的爱慕，他由此开始怀疑石神。出于对老友的理解，他决定独自调查此案，不为警方提供线索。最终，他识破了石神故布的疑阵，使石神无奈之下向警方自首。

石神真的是凶手吗？他是如何故布疑阵的？

要知道真相，就去读读原著吧，相信你会爱上这部经典的推理小说！

7.5　朴素贝叶斯：智能分类

“你好，我叫大白，你的私人健康助理。”

电影《超能陆战队》塑造了“史上最萌机器人”——大白，相信看过电影的朋友都想拥有一个像大白一样可爱的“私人健康助理”。所谓私人健康助理，是为个人进行健康服务的智能机器人，比如你突然发烧了，大白就会对你进行全身健康扫描，测试你的体温、白细胞数量等指标，它还会询问你的感受，然后对病情做出判断——肠炎（或病毒性感冒），给你对症下药，帮助你尽快恢复健康。我们不禁会好奇，大白怎么判断你得的是肠炎还是病毒性感冒呢？本节我们就来聊一聊大白看病的秘诀——朴素贝叶斯分类。

朴素贝叶斯分类是机器学习的一种方法，常用来解决分类问题。它是概率论在机器学习领域最重要的应用之一，其核心思想正是贝叶斯定理，也正是由于传承于形式简单的贝叶斯定理，我们才称为“朴素”贝叶斯分类。朴素贝叶斯分类常常应用于医学诊断，下面是一个典型案例。

疾病诊断

春天到了，北京街头飘起了杨絮，因为杨絮过敏而就医的人渐渐多了起

来，春夏之交，流感盛行，因为感冒而就医的人也渐渐多了起来。从症状来看，杨絮过敏和感冒十分相似，医生怎样判断病人是过敏还是感冒呢？

表 7-4 是某医院门诊近期的就诊情况记录，近期该医院门诊共接待了 20 位病人，症状有“打喷嚏”和“咳嗽”两种，男女病人数量相同，所患疾病有“感冒”和“过敏”两种。就在这时，又来了一位病人，性别女，症状是打喷嚏，她患感冒的概率是多少？

表 7-4　某医院门诊就诊记录

病人编号	症状	性别	疾病
1	打喷嚏	男	感冒
2	打喷嚏	男	感冒
3	打喷嚏	男	感冒
4	打喷嚏	男	感冒
5	打喷嚏	男	过敏
6	打喷嚏	男	过敏
7	咳嗽	男	感冒
8	咳嗽	男	感冒
9	咳嗽	男	感冒
10	咳嗽	男	过敏
11	打喷嚏	女	感冒
12	打喷嚏	女	过敏
13	打喷嚏	女	过敏
14	打喷嚏	女	过敏
15	打喷嚏	女	过敏
16	咳嗽	女	感冒
17	咳嗽	女	感冒
18	咳嗽	女	感冒
19	咳嗽	女	过敏
20	咳嗽	女	过敏

如果不使用贝叶斯定理，我们可能会这样计算：表中 11～15 号病人与新来的病人症状相同，这 5 位病人中有 1 位患有感冒，因此新来的病人患感冒的概率是 20％。

上述方法错在把“打喷嚏”和“性别女”作为一个条件来看待，它们本是两个彼此独立的条件，会各自独立地影响病人患感冒的概率，因此我们应当使用贝叶斯定理计算病人患感冒的概率。

由贝叶斯定理可得：

P(感冒 | 性别女且打喷嚏) = P(性别女且打喷嚏 | 感冒) · P(感冒) ÷ P(性别女且打喷嚏)

“性别女”和“打喷嚏”可以看作独立事件，因此：

P(感冒 | 性别女且打喷嚏) = P(性别女 | 感冒) · P(打喷嚏 | 感冒) × P(感冒)/[P(性别女) · P(打喷嚏)]

由表中数据可知：

P(感冒)=11/20；

P(性别女)=10/20；

P(打喷嚏)=11/20；

P(打喷嚏|感冒)=5/11；

P(性别女|感冒)=4/11。

将上面的数值代入贝叶斯定理的表达式，可以计算得到：

P(感冒 | 性别女且打喷嚏) = 33%

P(过敏 | 性别女且打喷嚏) = 67%

这便是使用朴素贝叶斯分类得到的诊断结果。

在实际应用中，医生掌握的病人信息会更多，医院的就诊记录也更多，但是朴素贝叶斯分类方法是不变的。

垃圾邮件识别

贝叶斯分类器的另一个典型应用是垃圾邮件识别。随着 E-mail 的普及，垃圾邮件也越来越猖獗。只要你的 E-mail 暴露于互联网上(比如用于账号注册)，便会迅速成为垃圾邮件的重灾区。垃圾邮件往往精于包装，配有令人诱惑的图片、词汇或附件，其中隐藏着很大的风险，比如盗号木马和网上诈骗。E-mail 用户厌恶垃圾邮件，但手动清理费时费力，还容易误点击，因此 E-mail 服务商很早就开始研究垃圾邮件的自动识别方法，最终他们选择贝叶斯分类器来识别垃圾邮件。

表 7-5 是一组垃圾邮件识别的基础数据，20 封邮件中有 10 封是垃圾邮件，10 封是普通邮件，用于判别的特征有三项——链接、图片和附件。第 21 封

邮件没有链接,但有图片和附件,它是垃圾邮件的概率为多少?

表 7-5　垃圾邮件识别的基础数据

邮件编号	链接	图片	附件	类别
1	有	有	有	垃圾邮件
2	有	有	有	垃圾邮件
3	有	有	没有	垃圾邮件
4	有	有	没有	垃圾邮件
5	有	没有	没有	垃圾邮件
6	有	没有	没有	垃圾邮件
7	有	没有	没有	垃圾邮件
8	没有	有	没有	垃圾邮件
9	没有	没有	有	垃圾邮件
10	没有	没有	没有	垃圾邮件
11	有	有	有	普通邮件
12	有	没有	没有	普通邮件
13	有	没有	没有	普通邮件
14	没有	有	有	普通邮件
15	没有	有	没有	普通邮件
16	没有	有	没有	普通邮件
17	没有	没有	有	普通邮件
18	没有	没有	没有	普通邮件
19	没有	没有	没有	普通邮件
20	没有	没有	没有	普通邮件

与疾病诊断不同,本例有三个特征,这不会影响贝叶斯定理的使用,只是计算方式上略有不同。根据贝叶斯定理,可知:

$$P(\text{垃圾邮件} \mid \text{无链接,有图,有附件}) = P(\text{无链接,有图,有附件} \mid \text{垃圾邮件}) \times P(\text{垃圾邮件}) / P(\text{无链接,有图,有附件})$$

$$P(\text{普通邮件} \mid \text{无链接,有图,有附件}) = P(\text{无链接,有图,有附件} \mid \text{普通邮件}) \times P(\text{普通邮件}) / P(\text{无链接,有图,有附件})$$

因为我们已知:

$$P(\text{垃圾邮件} \mid \text{无链接,有图,有附件}) + P(\text{普通邮件} \mid \text{无链接,有图,有附件}) = 1$$

因此我们只需要计算二者的比值,就可以计算出二者的数值。

先计算如下概率:

$$P_1 = P(\text{垃圾邮件}) = 5/10$$

$$P_2 = P(\text{普通邮件}) = 5/10$$

$$P_3 = P(\text{无链接} \mid \text{垃圾邮件}) = 3/10$$

$$P_4 = P(\text{无链接} \mid \text{普通邮件}) = 7/10$$

$$P_5 = P(\text{有图} \mid \text{垃圾邮件}) = 5/10$$

$$P_6 = P(\text{有图} \mid \text{普通邮件}) = 4/10$$

$$P_7 = P(\text{有附件} \mid \text{垃圾邮件}) = 3/10$$

$$P_8 = P(\text{有附件} \mid \text{普通邮件}) = 3/10$$

再计算所求两个概率的比值：

$P(\text{垃圾邮件} \mid \text{无链接,有图,有附件})/P(\text{普通邮件} \mid \text{无链接,有图,有附件})$

$=P(\text{无链接,有图,有附件} \mid \text{垃圾邮件}) \cdot P(\text{垃圾邮件}) \div$

$P(\text{无链接,有图,有附件} \mid \text{普通邮件}) \cdot P(\text{普通邮件})$

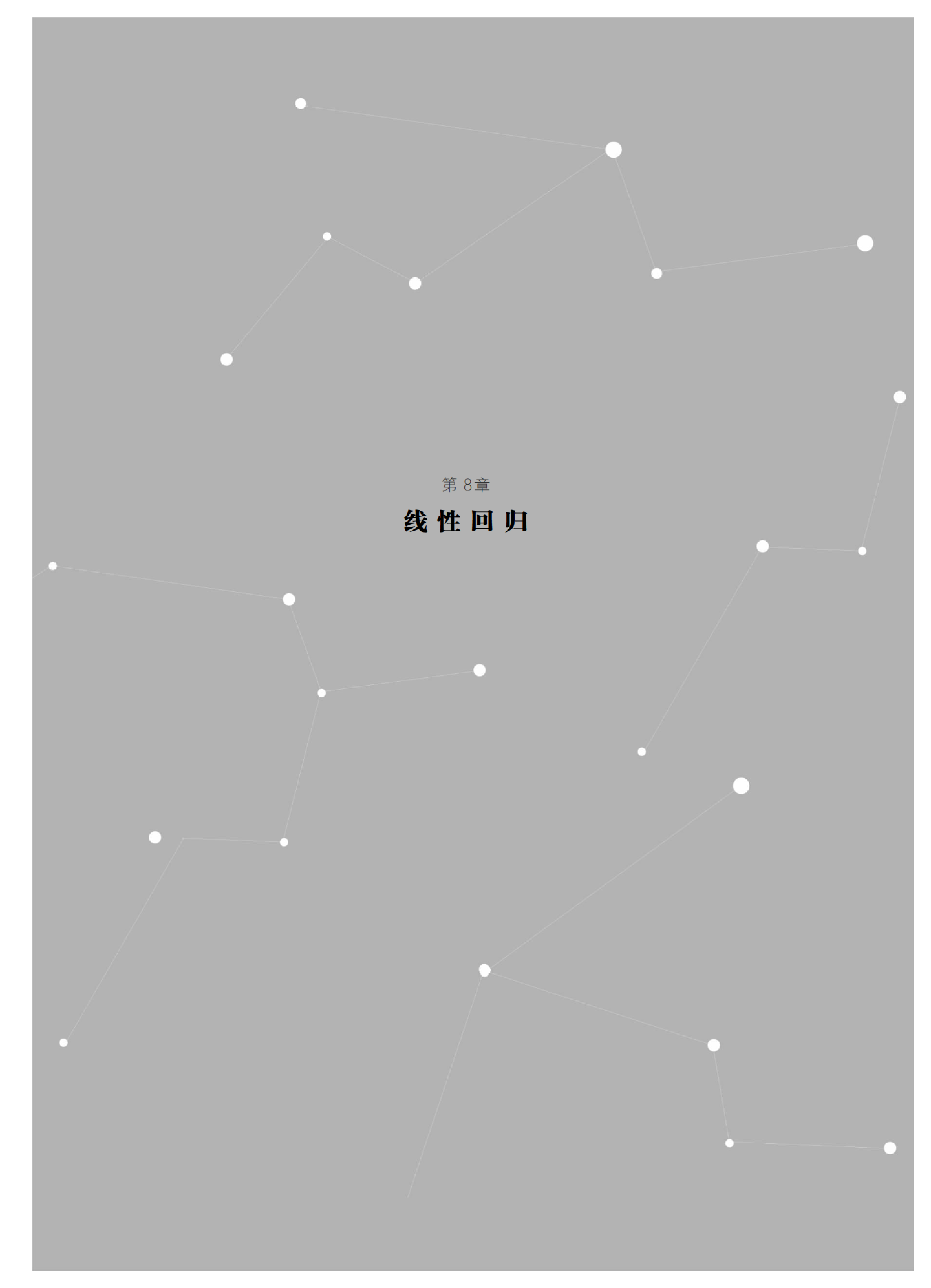
$=P_1 \cdot P_3 \cdot P_5 \cdot P_7/(P_2 \cdot P_4 \cdot P_6 \cdot P_8)$

$=15/28$

因此，$P(\text{垃圾邮件}|\text{无链接,有图,有附件})=15/43=35\%$

也就是说，一封无链接、有图、有附件的邮件是垃圾邮件的概率是35%，是普通邮件的概率是65%。

最后需要说明的是，朴素贝叶斯分类器包含一个关键假设：各个特征互相独立。这个假设在大多数实际问题中都是成立的，但是我们不能因此忽略这个假设。



第 8章

线 性 回 归

导语：2013 年 8 月，谷歌公司提出了一个票房预测模型，该模型仅以单词搜索量为依据，便可以提前一个月预测电影的首周票房，准确度高达 94%。更令人惊讶的是，这是一个简单的线性回归模型。谷歌是如何做到的？

8.1 预测未来：以数据之名

凯文·凯利（Kevin Kelly，绰号 KK）是个难以定位的人物，他曾是科技杂志 *Wired* 的主编，他是周游世界的游侠，他还是一位科技哲学家，曾撰写多部科技哲学著作。KK 的第一部"神作"是 1994 年出版的《失控》，这部书不仅揭示了网络文化的内涵，甚至预言了网络文化的兴起。当时这部书读起来像一部长篇科幻小说，但互联网摧枯拉朽般地发展印证了书中所写。从这一点来看，KK 更像是一个科技预言家，他早于世人看清了网络文化的本质，预言了网络文化的盛行。

在凯文·凯利的新书《必然》中有这样一段描述：

2002年左右，我参加了一家小公司举办的聚会，其间，我问这家公司的创始人拉里·佩奇："拉里，我搞不懂，已经有这么多家搜索公司了，你们为什么还要做免费网络搜索？"拉里·佩奇回答说："哦，我们其实是在做人工智能。"

拉里·佩齐正是谷歌公司的创始人。谷歌公司在新千年伊始就瞄准了人工智能技术，这同样是一次大胆的预言，事实证明，这个预言应验了。在过去的十几年里，谷歌收购了多达13家人工智能和机器人公司，制作出了安卓手机系统、谷歌地图、谷歌眼镜、无人驾驶汽车、无人机等多款智能产品。在谷歌看来，人工智能并非是机器人代替人类来工作，人工智能要做到人类做不到的事——预测未来。

谷歌流感趋势

2008年年初，谷歌推出了"谷歌流感趋势"(Google Flu Trends，GFT，网址 https://www.google.org/flutrends)，这个工具根据谷歌搜索数据的汇总，近乎实时地对全球当前的流感疫情进行估测。当时，"大数据"的概念尚未普及，数据预测技术还处于萌芽期，GFT并未引起广泛关注。2009年，谷歌使用GFT不仅成功预测到H1N1在全美范围的传播，而且对病毒爆发时间和地点判断极其准确，媒体纷纷报道了这次令人称奇的预测，GFT引起了全世界的关注。与习惯性滞后的官方数据相比，谷歌成为一个更有效、更及时的预测指标。

其实，谷歌的工程师们很早就发现：在流感季节，与流感有关的搜索量会明显增多；到了过敏季节，与过敏有关的搜索量会显著上升；而到了夏季，与晒伤有关的搜索量又会大幅增加。我们知道，没有任何患病症状的人是不会去搜索疾病相关的关键词的，因此，疾病相关的关键词搜索量很可能有助于了解疾病的传播和分布情况。2009年2月的 *Nature* 杂志刊发了一篇题为 *Detecting influenza epidemics using search engine query data* 的论文，文中介绍了GFT的原理。谷歌以相关性为衡量指标，找到了45个与流感就诊密切相关的搜索关键词，然后以这45个关键词的搜索量为参考值，估算流感症

状的就诊比例。图 8-1 是预测结果与实际数据的对比图，超前两周的曲线表示预测结果随时间的变化，滞后两周的曲线表示实际就诊比例随时间的变化，两条曲线一直十分接近，说明预测得非常准确。

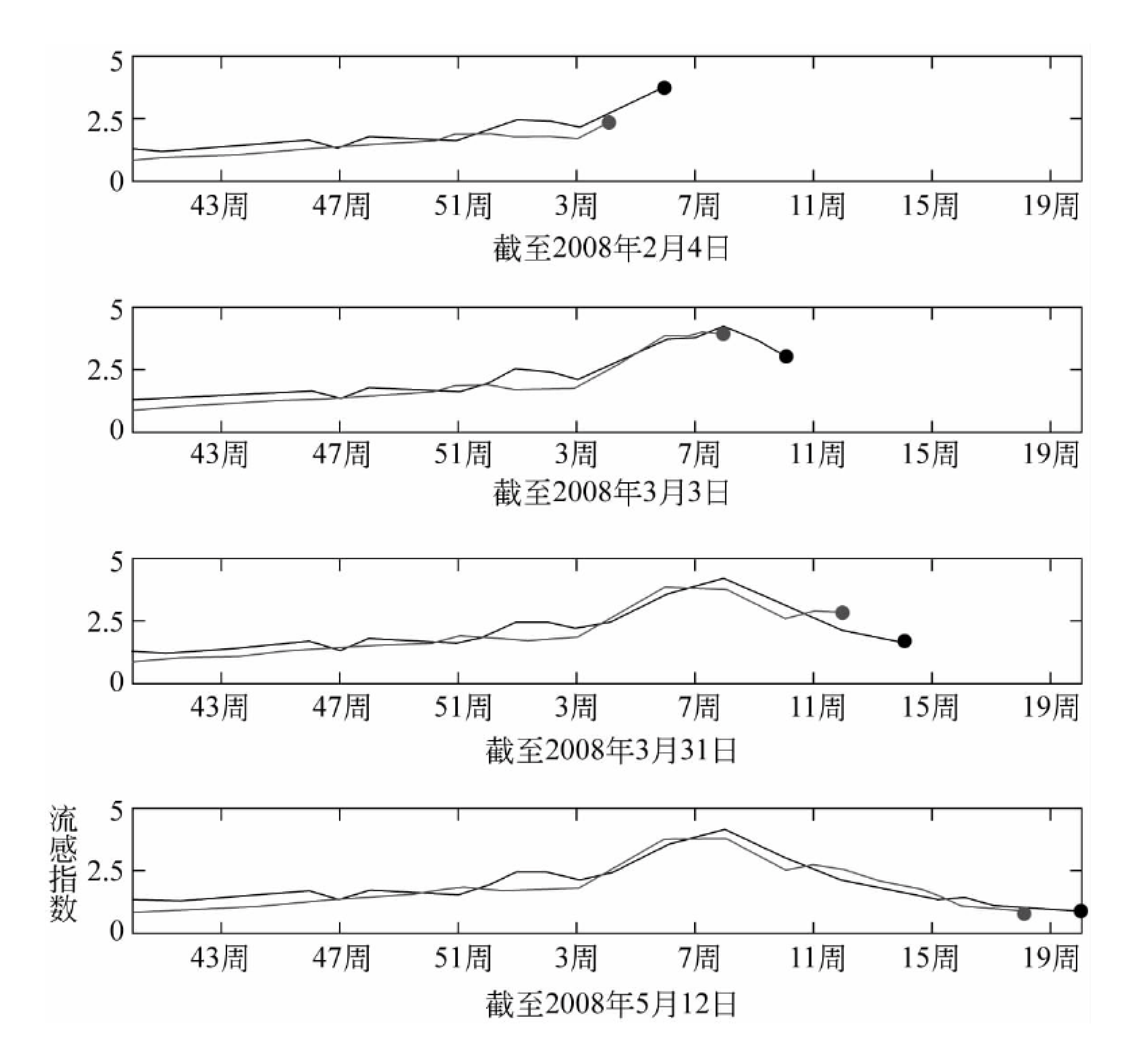


图 8-1　GFT 的预测结果与实际数据的对比

然而，GFT 在受到世界瞩目之后，却遭遇了尴尬的“见光死”。2013 年 1 月，季节性流感再次在美国爆发，这一次 GFT 遭遇了“滑铁卢”，它预测的就诊数据比实际数据高出两倍之多。媒体报道了 GFT 的错误预测，并且指出，在 2013 年之前的很长一段时间内，GFT 都高估了流感疫情。从 2011 年 8 月—2013 年 9 月的108 周中，GFT 高估流感疫情长达 100 周。这些错误不是随机分布的，说明 GFT 的确出现了错误。

从精准的预测，到巨大的错误，GFT 的大起大落令人唏嘘。但不可否认的是，GFT 是一次伟大的尝试，是数据预测技术的一次零的突破，从此数据预测渐渐成为科技领域的热门课题。

预测世界杯

随着大数据概念的兴起，众多科技巨头开始钻研数据预测技术。在体育、娱乐等领域做预测格外受到青睐，一方面可以检验算法，另一方面还可以借助广泛的球迷、影迷基础做一次免费广告。于是，2014 年巴西世界杯成为科技巨头展示数据预测技术的舞台。

这一次不再是谷歌的独角戏，微软、高盛和中国的百度与谷歌一同玩起了“大数据预测世界杯”的游戏。2014 年 6 月 12 日，世界杯小组赛正式开始，百度、微软和高盛对 48 场小组赛进行了预测，百度以 58％的准确率领跑，微软和高盛分别以 56.25％和 37.5％的准确率排在第二、第三位。此后，四家公司全部参与了淘汰赛阶段的预测，百度和微软预测正确了全部 16 场淘汰赛的胜负结果，以 100％的预测准确率震惊了全世界！谷歌错误地预测了法国队会战胜德国队，遗憾未能实现 100％的预测准确率。

世界杯后，媒体披露了四家公司各自的预测方法。百度以过去五年国际赛事数据和 400 多家博彩公司的赔率为参考数据，计算球队实力、近期状态、主场效应、博彩数据和大赛能力五项指标，采用多源数据融合技术进行预测；谷歌则只以 Opta Sports 网站的比赛数据为参考数据，计算各球队和球员的技战术能力指标，然后采用计算机排序算法进行预测。预测错误之后，谷歌官方博客称，德国队和法国队的比赛预测失败的最重要原因是，赛事数据量过大以及球员跑动射门等指标的错误计算。

仅靠一次世界杯的预测结果，并不能说明哪一种数据预测方法更有效。时至今日，数据预测仍然是一门新兴技术，概率统计、机器学习、深度学习甚至数据融合都可以应用到数据预测中。接下来，我们就来学习概率统计中的数据预测技术——回归分析。

8.2 线性回归：奇准的票房预测

2013 年 8 月，谷歌公司把大数据技术成功应用到电影票房的预测上，并撰

文公布了研究成果 *Quantifying Movie Magic with Google Search*。该报告称,谷歌的预测模型可以提前一个月预测电影上映的首周票房,准确度高达 94%。令人吃惊的是,谷歌并没有搜集各种电影相关的数据来提高预测准确度,而是仅仅使用了他们自有的数据——单词搜索量,而且,谷歌的预测模型居然是概率统计中最简单的线性回归模型。

据谷歌统计,从 2011—2012 年,谷歌的电影相关搜索量增长了 56%,正是由于人们越来越多地使用谷歌搜索电影相关信息,才使得谷歌萌发了票房预测的想法。谷歌的工程师们画出了 2012 年电影相关的搜索总量和票房总收入的曲线图,如图 8-2 所示,实线表示电影相关关键词的搜索量随时间的变化趋势,虚线表示电影票房随时间的变化趋势,两条曲线的起伏变化十分相似。

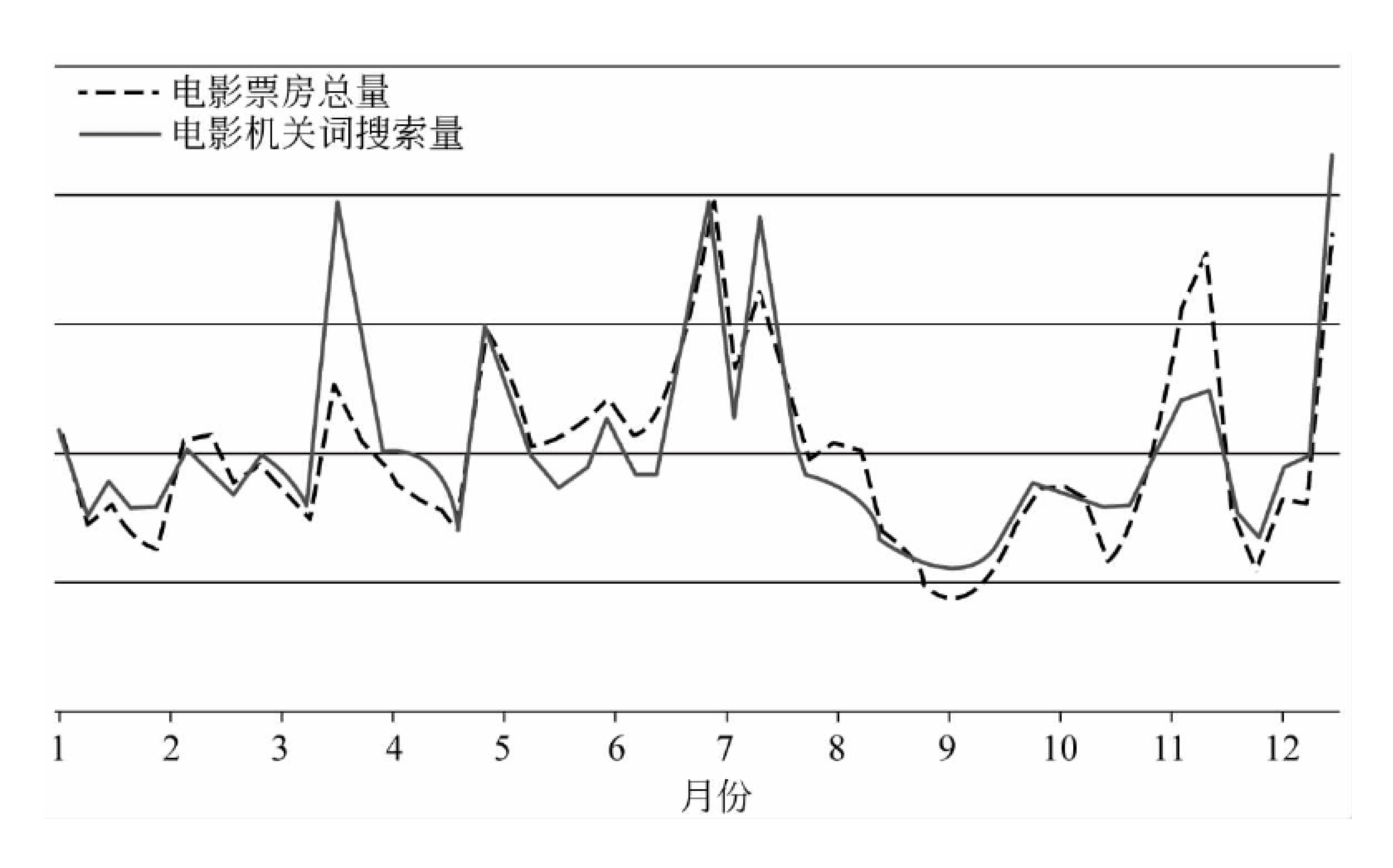


图 8-2　2012 年电影票房和电影相关词搜索量随时间的变化曲线

如此相似的两条曲线激起了谷歌工程师的好奇心,这似乎预示着两条曲线存在很强的相关性。谷歌的工程师们将电影搜索进而分为两类——电影名搜索和电影相关词搜索,并画出两类搜索量和票房收入的关系。如图 8-3 所示,虚线仍然表示电影票房随时间的变化趋势,起伏较大的实线表示电影名搜索量随时间的变化趋势,较平坦的实线表示电影相关词的搜索量随时间的变化趋势。图 8-3 中曲线显示,电影名往往比电影相关词的搜索量更大,但在电影上映的淡季(图 8-3 中阴影部分),电影相关词的搜索量反超了电影名的搜索量,这是因为那时没有好看的电影,人们会转而搜索诸如“好莱坞电影”“功

夫片”之类的词汇。两类关键词搜索量的变化趋势与票房变化趋势仍然十分相似。

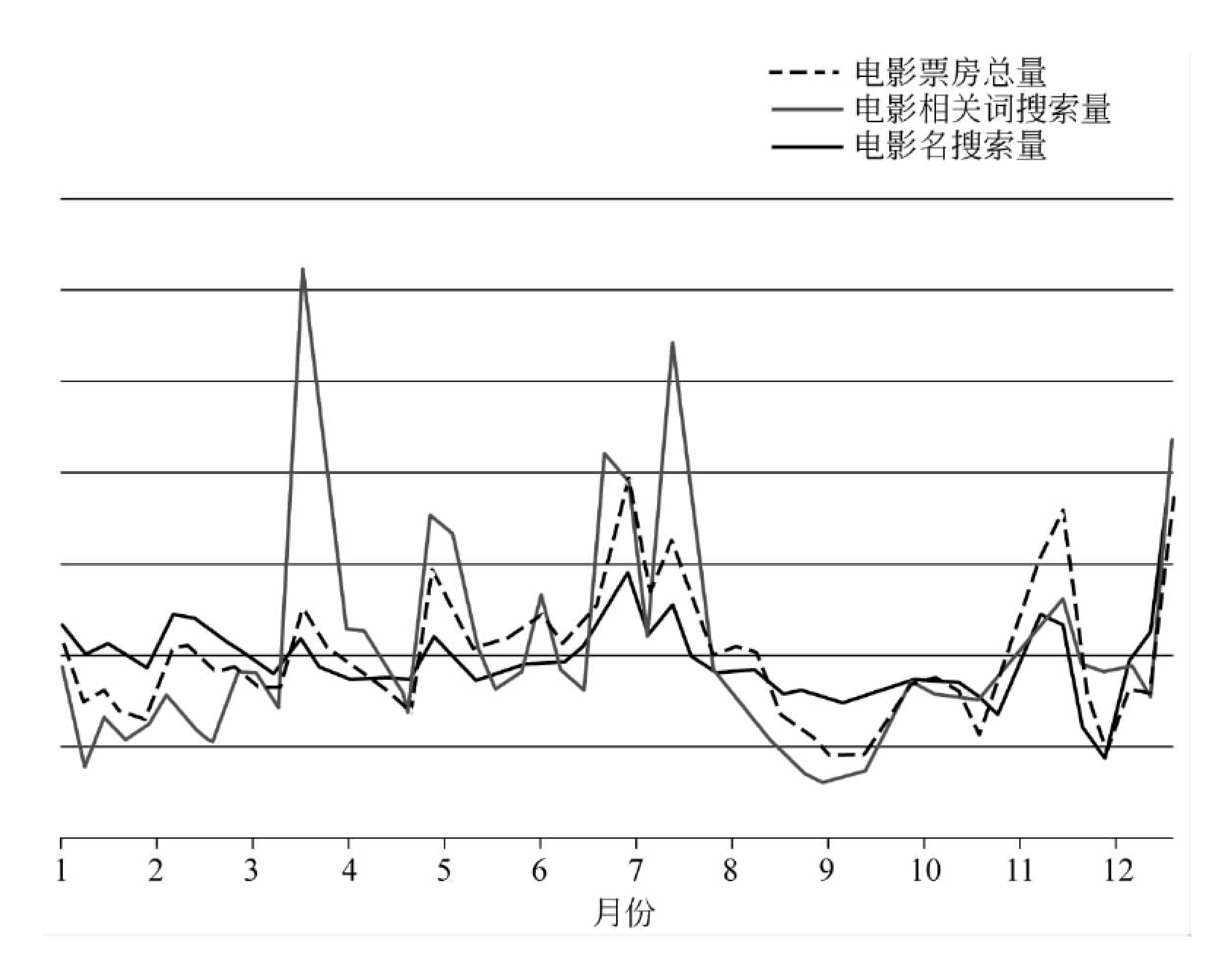


图 8-3　2012 年电影票房和两类关键词搜索量随时间的变化曲线

前面的研究似乎说明了搜索量和票房之间强烈的相关关系，所以，谷歌要再进一步：提前一周预测一部电影的票房。谷歌选取了 2012 年上映的 99 部电影，画出了搜索量和票房的关系图，并试图构建一个线性模型，可是预测准确度只有 70％，如图 8-4 所示。为了提高预测准确度，谷歌需要搜集更多的数据，经过反复的试验，它们选定了放映前一周的搜索量、广告点击量、上映影院数量和同系列电影前几部的票房表现四类指标，重新构建线性模型，将预测准确率一举提高到了 92％。

可惜的是，提前一周预测票房对电影的营销几乎没有帮助，因为在电影上映前一周，营销策略几乎无法更改，即使更改，效果也来不及体现。因此，谷歌需要挑战更高的难度——提前一个月预测。

在电影上映前一个月，电影的搜索量还不够多，难以用来预测，谷歌挖掘出了另一个更有说服力的指标——电影预告片的搜索量。现在，几乎每部电影都会在放映前投放预告片，观众也喜欢在影片上映前搜索预告片来观看，因此，谷歌将预告片的搜索量作为票房预测的一个指标。除此之外，谷歌还选择

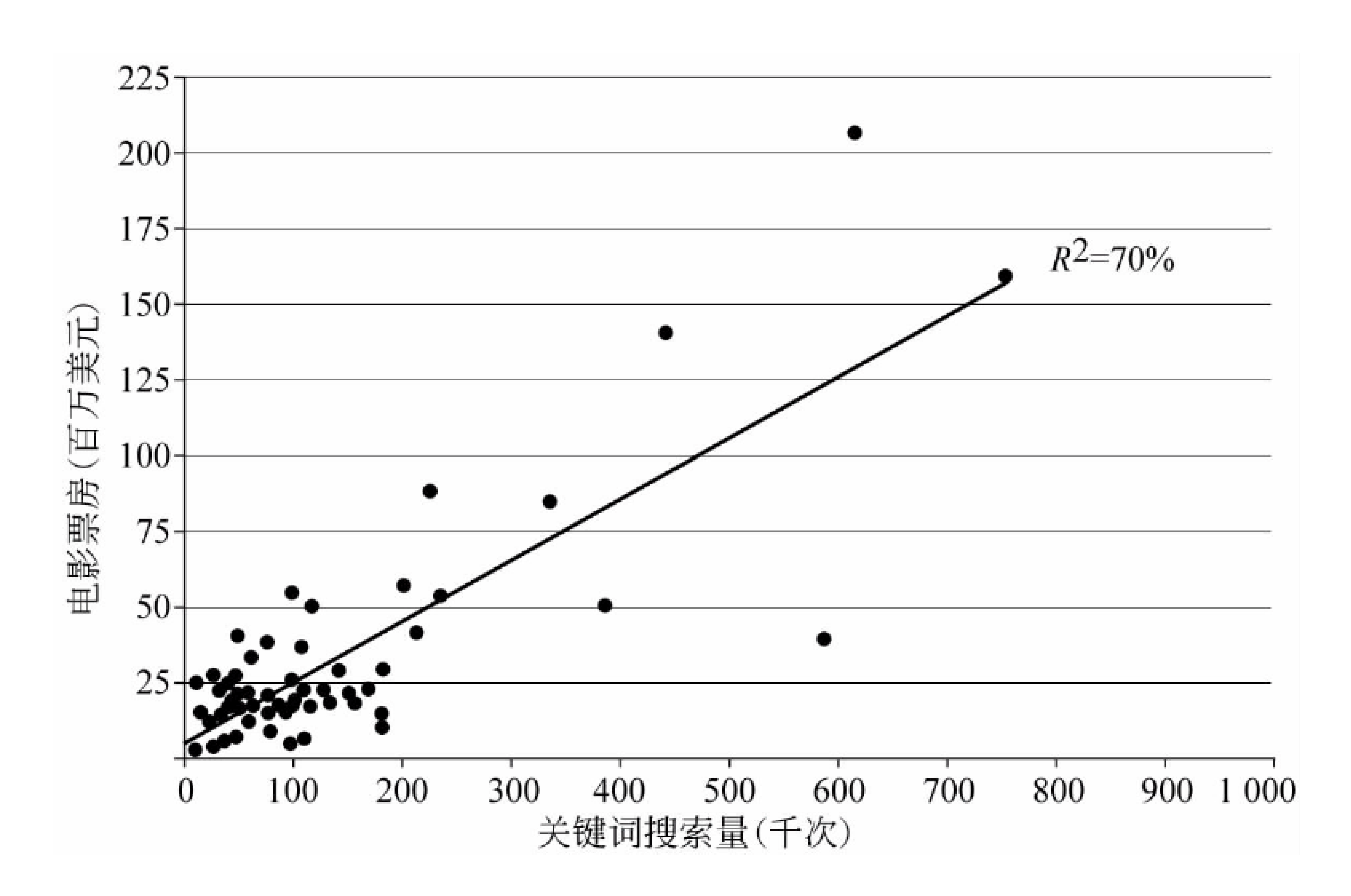


图 8-4　99 部电影的票房和搜索量的线性回归模型

了以同系列电影前几部的票房和档期的旺季淡季特征作为参考指标，使用这些指标构建的线性模型最终实现了准确率高达 94%的预测。

线性回归

回归分析是一种统计分析方法，用于研究多个统计量之间的关系，并利用关系进行预测。线性回归模型是最简单的回归分析模型，下面我们尝试复盘谷歌的分析过程，应用线性回归来预测票房。

图 8-5 是计算机模拟生成的 500 个数据点，每个点表示一部电影，横坐标是预告片搜索量，纵坐标是票房。图 8-5 称为散点图，是统计分析中最简单、最常用的图，用于对数据的规律做初步观察。观察图 8-5 可以发现，这些数据点大多分布在一条直线附近，这条直线代表了这些数据的分布规律，线性回归要做的就是根据散点图找到这条直线，这一过程也称为线性拟合。

设拟合直线的方程是 $y=ax+b$，x 表示预告片搜索量，y 表示首周票房。线性回归的目标是找到最能体现数据特征的直线，也就是说，这条直线需要尽可能地“接近”所有数据。衡量多个点和一条直线之间的“接近程度”，最常用的指标是误差平方和。图 8-6 是误差平方和的一个示意图，基础数据包含 4 个点(图中的空心圆圈)，这四个点的 X 坐标分别对应拟合直线上的四个 Y 坐

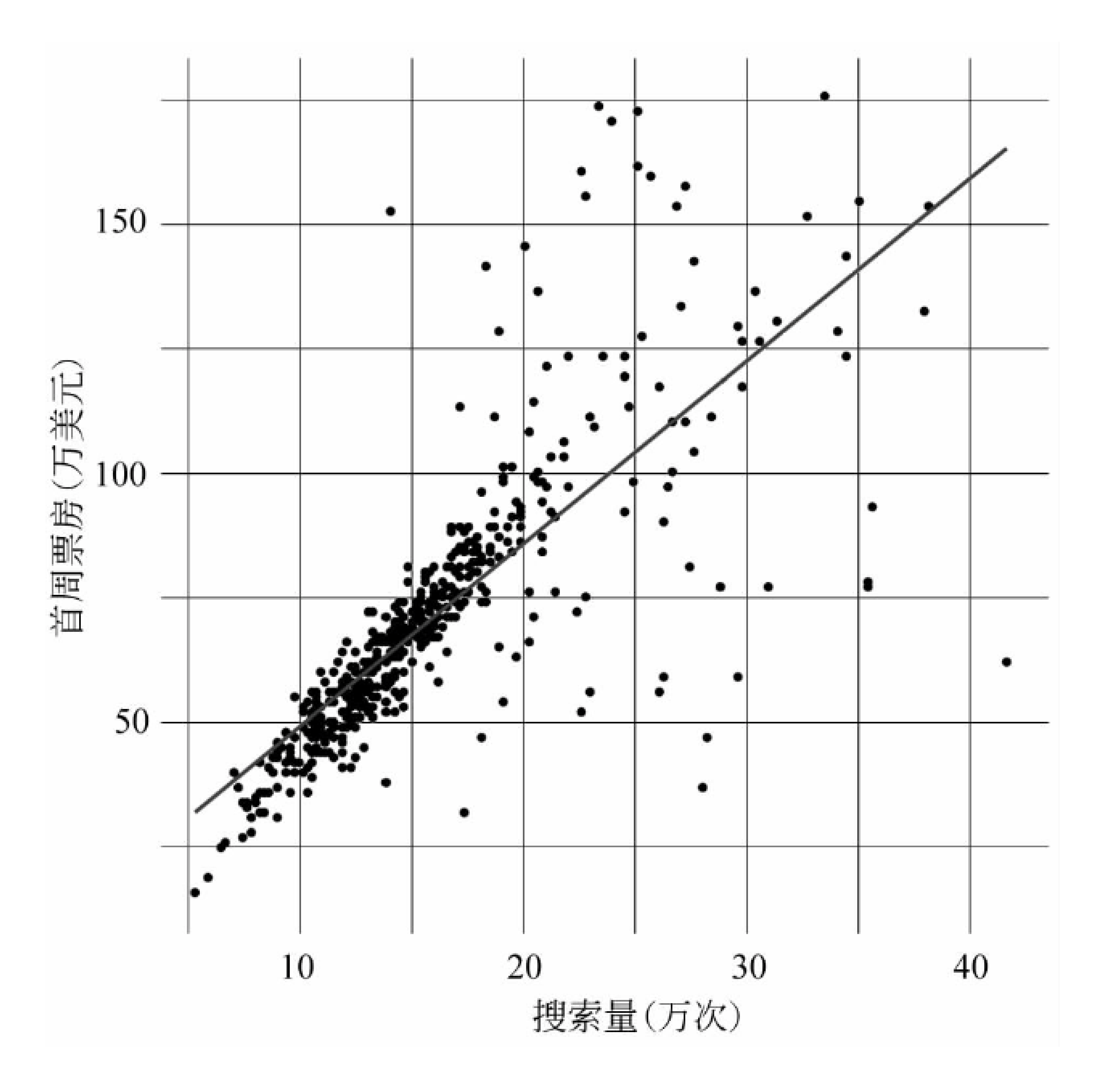


图 8-5　首周票房和预告片搜索量的散点图

标,图中四条虚线的长度的平方和就是误差平方和,使误差平方和最小的那条直线就是最佳拟合直线,这种求解方法也称为最小二乘回归法。

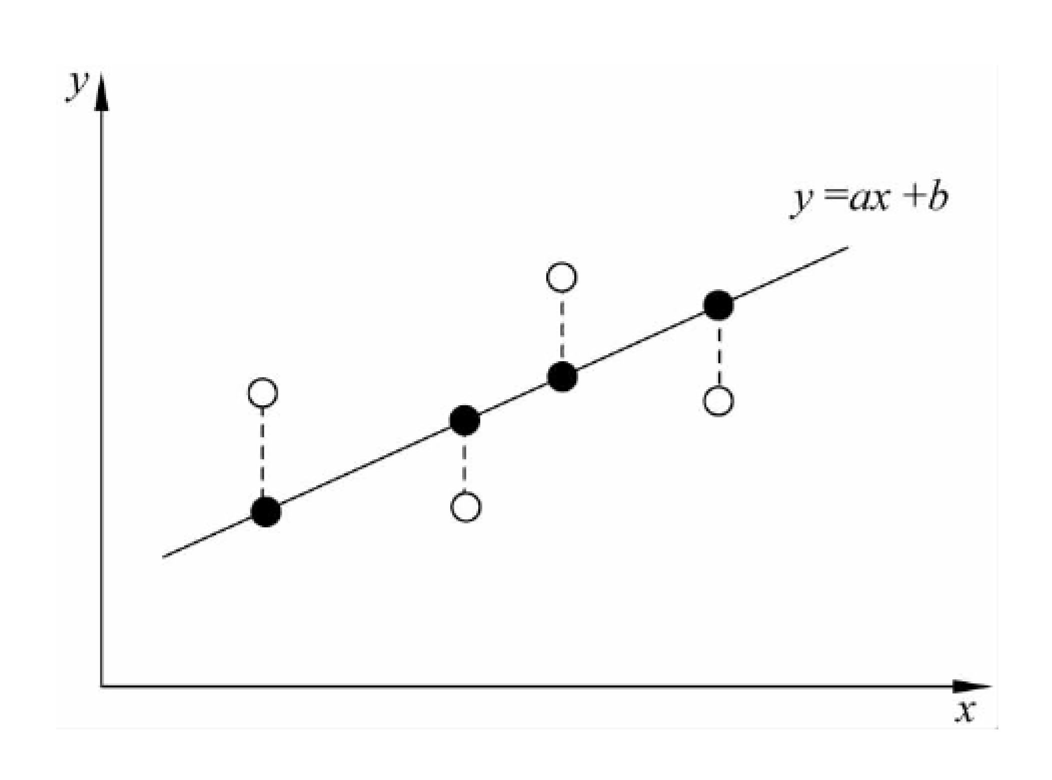


图 8-6　误差平方和示意图

当误差平方和达到最小值时,可以计算出 a 和 b 的值为

$$a=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}$$

$$b=\overline{Y}-a\overline{X}$$

至此便计算出了最佳拟合直线的表达式。

在处理线性回归问题时，我们可以把数据代入公式中进行计算，也可以使用统计软件，如 Excel、R、SPSS 等常用统计软件都有线性回归函数，我们只需要做少量的操作或编码就可以计算出线性回归的结果。

经计算，票房和搜索量的线性回归直线方程是：

$$y = 3.5x + 13.6$$

这条直线代表了票房和搜索量之间的关系，如图 8-7 所示。我们可以使用这条直线来预测票房，比如，某部即将上映的影片，预告片搜索量是 12 万次，即 $x=12$，根据直线方程可以计算出 $y=55.6$，因此我们预测这部影片的首周票房是55.6 万美元。

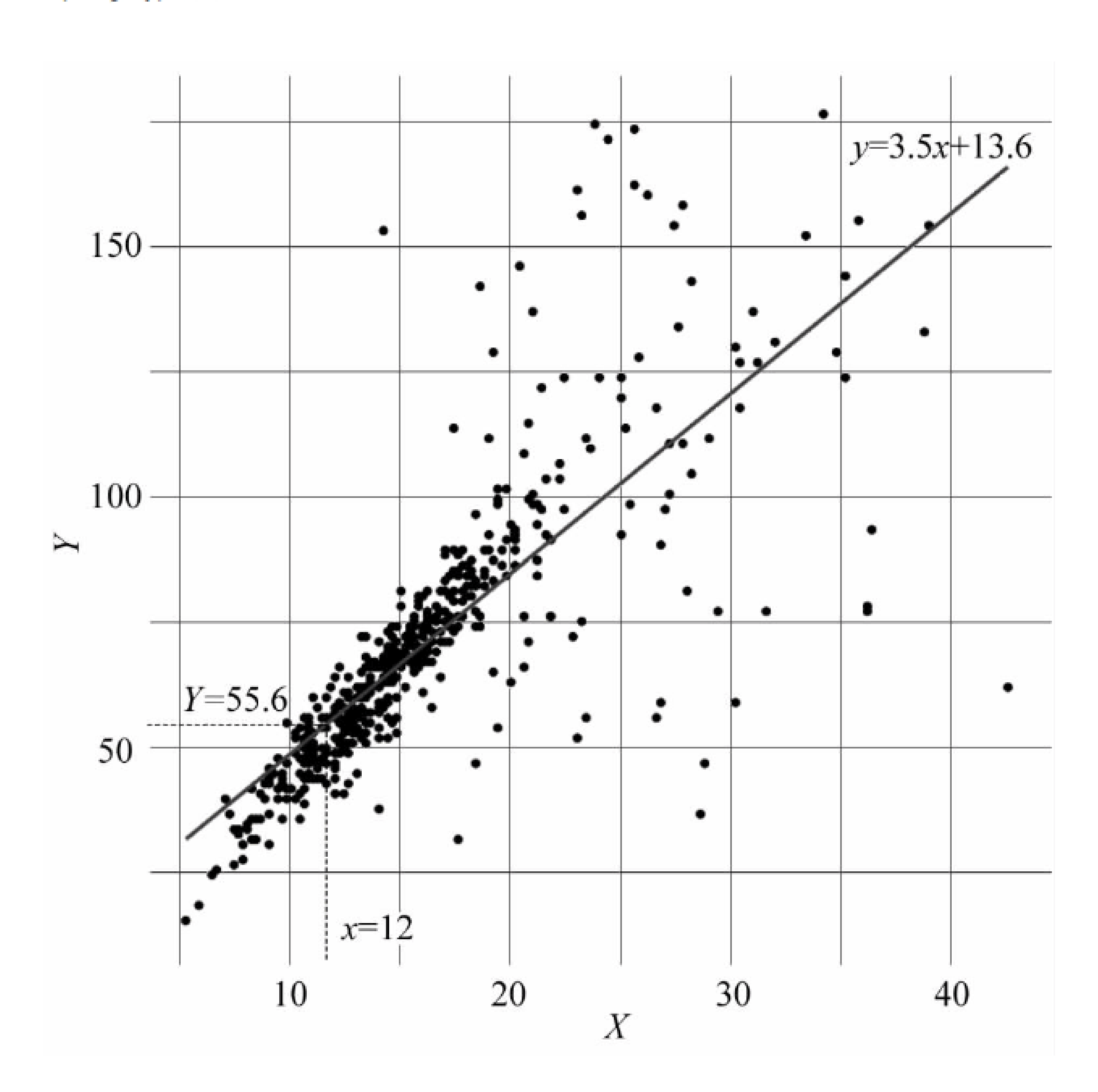


图 8-7　线性回归结果

除了直线方程，我们还可以计算另一个量化指标——相关系数。相关系数可以帮助我们判断两个变量的线性相关关系。此前，我们观察散点图，已经发现票房和搜索量之间近似存在线性相关关系，这只是感性判断，相关系数是对线性相关关系的理性判断。

相关系数 r 的计算公式为

$$r = a \times S_x / S_y$$

式中，a 是直线方程中的 a，S_x 表示 X 的标准差，S_y 表示 Y 的标准差。如图 8-8 所示，r 可以是 -1～1 的任意数值，其中最特别的三个数值是 -1、1 和 0，含义如下：

$r=-1$ 表示 y 和 x 存在负相关关系，即 a 是负数；

$r=1$ 表示 y 和 x 存在正相关关系，即 a 是正数；

$r=0$ 表示 y 和 x 不存在任何线性相关关系，即 $a=0$，不存在拟合直线。

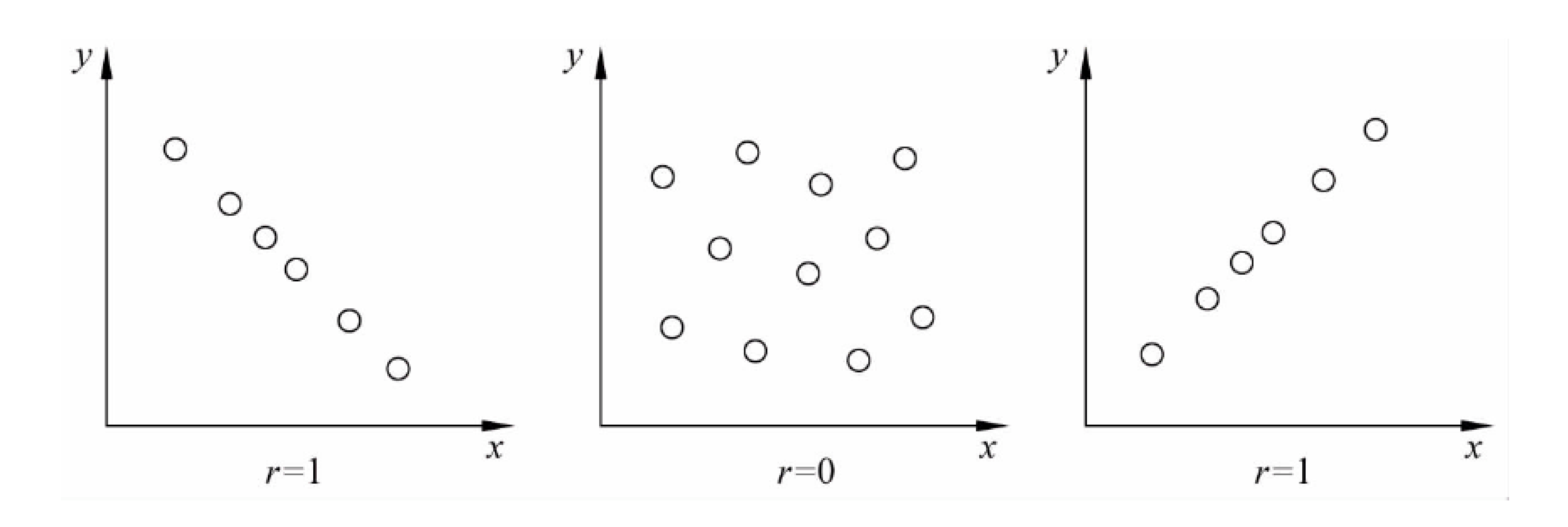


图 8-8　线性相关系数 r 的示意图

在实际问题中，r 的值大多不会是 -1、1 或 0，但我们可以借助它们的含义来判断线性相关关系。比如，当 $r=0.9$ 时，我们认为 r 的值接近 1，y 和 x 存在近似的正相关关系；当 $r=-0.9$ 时，我们认为 r 的值接近 -1，y 和 x 存在近似的负相关关系；当 $r=0.05$ 时，我们认为 r 的值接近 0，y 和 x 几乎不存在线性相关关系。

至此，我们计算出了线性回归方程和线性相关系数，这只是线性回归分析的第一步。接下来我们还要对线性回归的结果进行评估和改进。

8.3　拟合评估：拟合优度与分区段拟合

拟合优度

谷歌曾在电影票房预测模型中提到，它的预测可以达到 94%的准确率，如图 8-9 所示。这里提到的“94%准确率”很容易被误解为，平均 100 部影片有

94 部能预测正确，或者预测结果与实际票房相差 6%。这两种理解都不对。94%代表的是线性回归模型的拟合优度。

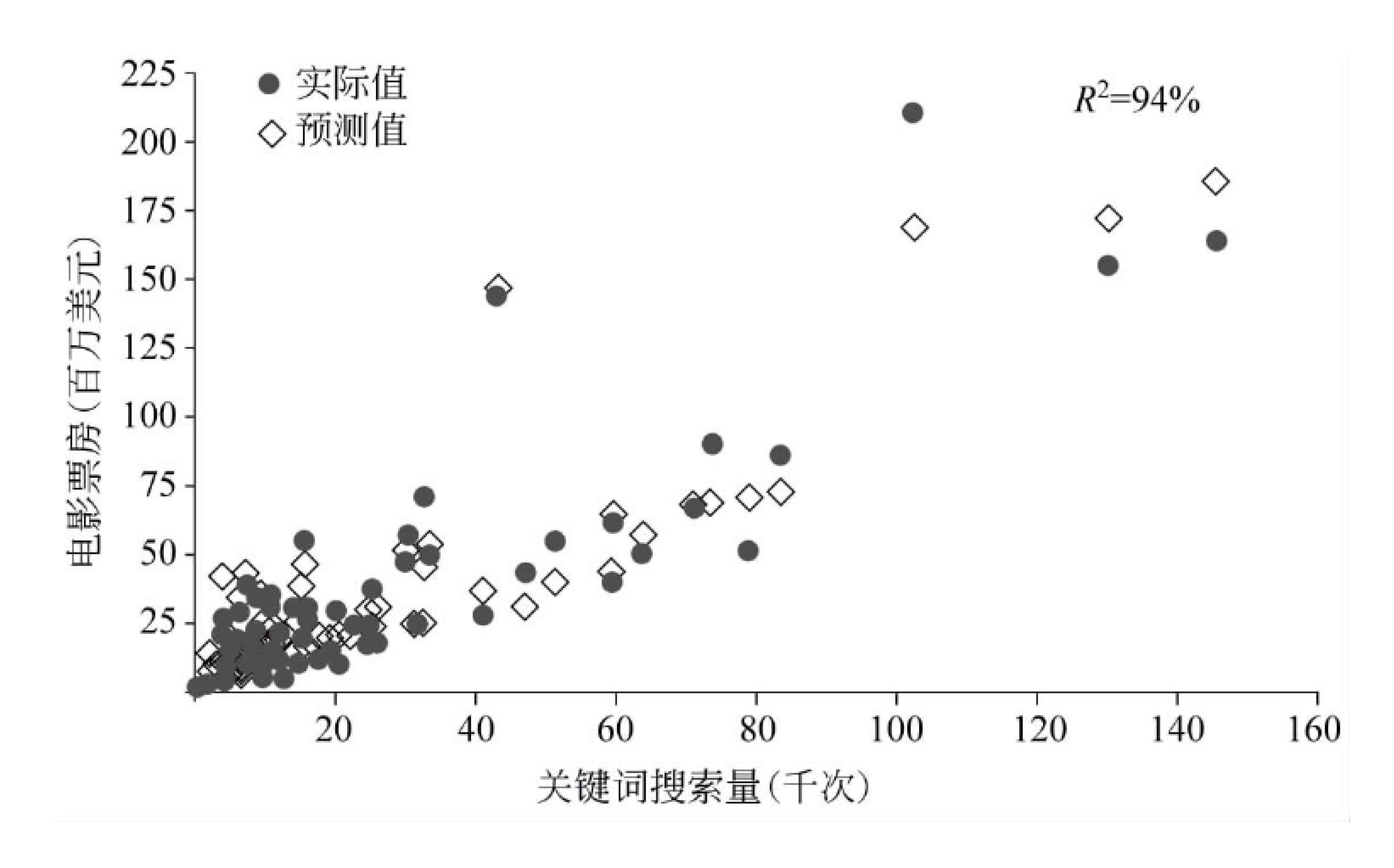


图 8-9　谷歌票房预测模型可达到 94%的准确率

拟合优度，亦称决定系数、判定系数，是用于评价线性回归模型有效性的指标，记为 R^2。拟合优度的取值在 0～1，越接近 1，模型越有效，越接近 0，模型越无效。拟合优度的计算公式为

$$R^2 = SSR/SST = 1 - SSE/SST$$

其中，SST(Sum of Squares for Total，SST)表示总平方和，SSR(Sum of Squares for Regression，SSR)表示回归平方和，SSE(Sum of Squares for Error，SSE)表示误差平方和，三者之间的关系是 $SST=SSR+SSE$，三者的计算公式为

$$SST = \sum (y_n - \bar{y})^2$$

$$SSR = \sum (\hat{y}_n - \bar{y})^2$$

$$SSE = \sum (\hat{y}_n - y_n)^2$$

其中，y_n 表示第 n 个样本，$\hat{y}_n$ 表示第 n 个样本的预测值，$\bar{y}$ 表示样本均值。

将上一节的基础数据做线性回归，可以得到 $R^2=61.4\%$，这就是回归直线 $y=3.5x+13.6$ 对应的拟合优度。

在上一节中，我们曾提到过误差平方和 SSE，根据误差平方和的定义，它可以用来衡量拟合效果，为什么不用 SSE 而要用 R^2 呢？因为 SSE 不具备可

比性，R^2 具备可比性。SSE 是一个绝对数值，对于同样一组数据，不同的拟合结果之间可以用 SSE 来对比，SSE 越小，拟合效果越好。可是，在实际问题中，数据常常是动态变化的，不同的数据得到的拟合结果，无法用 SSE 来对比，因为 SSE 与数据量有关。R^2 是一个相对数值，它有明确的取值范围，取值的边界也有明确的意义，不同的数据计算出的 R^2 与数据量无关，因此不同拟合结果的 R^2 可以进行对比，R^2 越接近 1，拟合效果越好。谷歌票房预测模型的拟合优度达到 94%，十分接近 1，说明拟合效果非常好。

分区段拟合

线性回归也有自己的局限性。观察图 8-10 可以发现，所有已知电影的搜索量都分布在 5 万～43 万次这个区间内，这说明拟合得到的直线只能用于预测这个区间内的电影票房，如果某部电影的预告片搜索量是 4 万次或 44 万次，拟合结果将无法做出准确的预测。此外，在 5 万～43 万次这个区间里，预测效果也是有区别的，当搜索量处于 5 万～20 万次时，拟合直线与基础数据更接近，预测效果也越好，当搜索量处于 20 万～43 万次时，拟合直线与基础数据

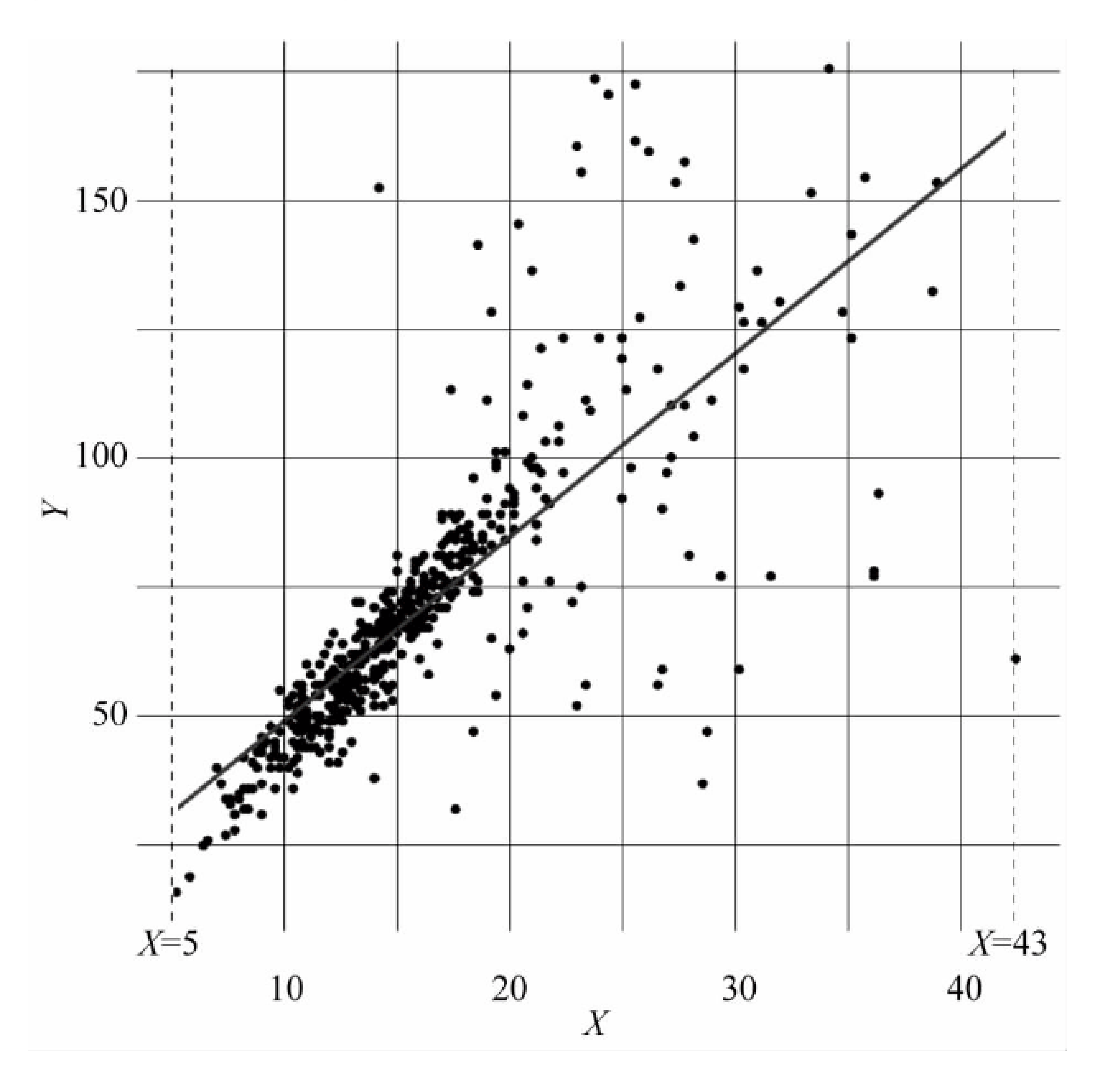


图 8-10　对线性回归结果的再观察（万次）

相距更远，预测效果并不好。这提示我们，可以将这两个区间分别做线性回归，这就是分区段拟合。

我们将基础数据分为 $X\leqslant 20$ 和 $X>20$ 两部分，分别进行线性回归，可以得到两条回归直线，如图 8-11 所示。

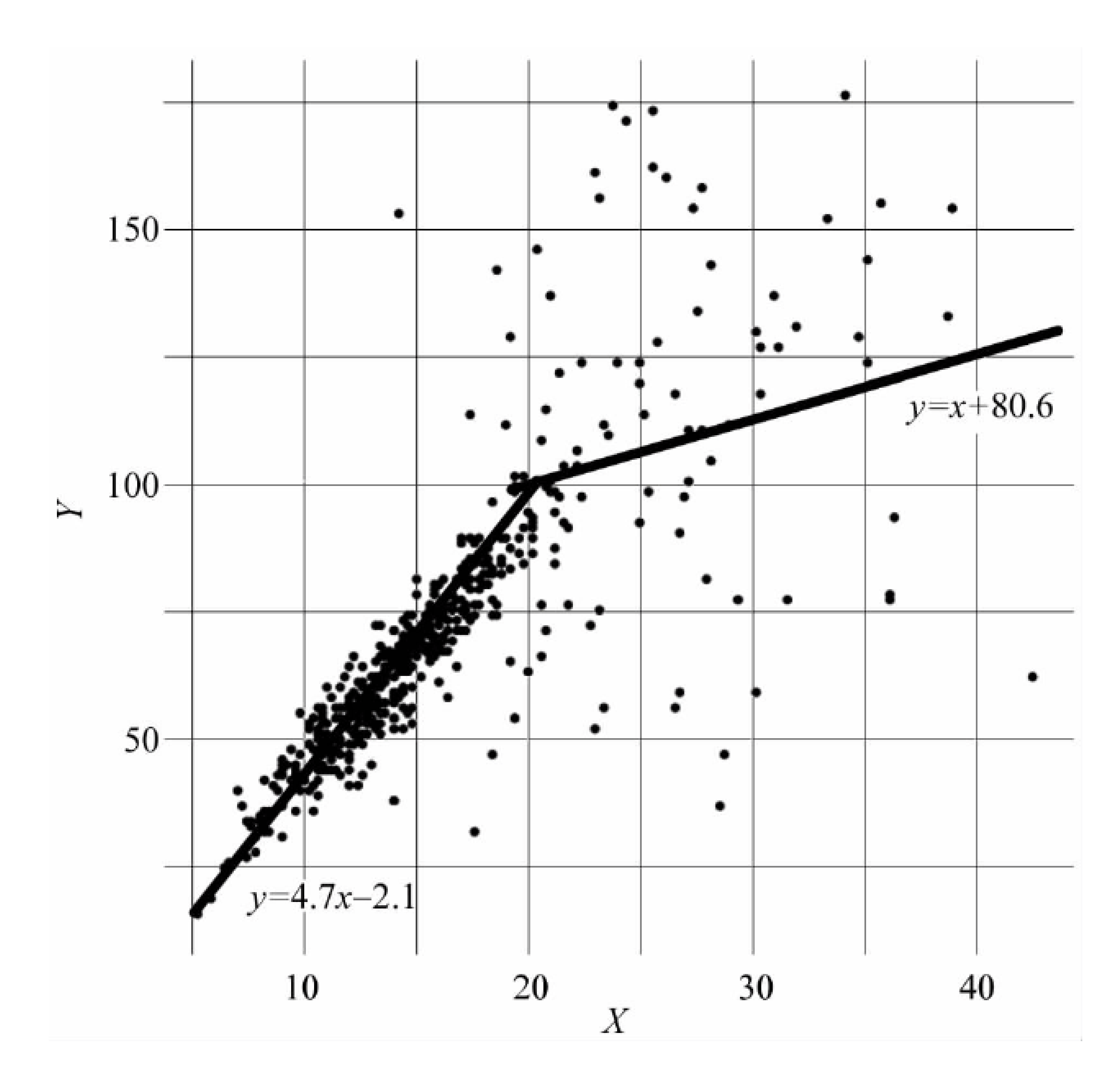


图 8-11　分区段拟合结果（万次）

当 $X\leqslant 20$ 时，回归直线方程为：$y=4.7x-2.1$，拟合优度为 72.6%。

当 $X>20$ 时，回归直线方程为：$y=x+80.6$，拟合优度为 3%。

相比于只做一次线性回归，分区段拟合提高了 $X\leqslant 20$ 时的拟合优度，$X\leqslant 20$ 时回归模型对 Y 值的预测会更准确。$X>20$ 时，拟合优度只有 3%，说明拟合效果较差，这个区间的预测准确率也会比较低。

过拟合

既然可以分两个区段拟合，能否分三个、四个、五个甚至十个区段呢？是不是区段划分得越多，模型的拟合效果越好呢？答案是否定的，因为存在过拟合现象。

过拟合，顾名思义，是指过度拟合，图 8-12 是过拟合的典型案例。基础数

据被强行划分为多个区间，分别进行线性回归，得到多条回归直线。这样的划分看似精益求精，却违背了线性回归的核心思想：寻找数据的隐含规律。回归直线并不是要把已知数据连接起来，而是从全局的角度描述数据的隐含特征，数据并不需要全部落在回归直线上，因为误差总是存在的。“过拟合”没有公认的判断标准，只能靠我们在实践中学习体会。

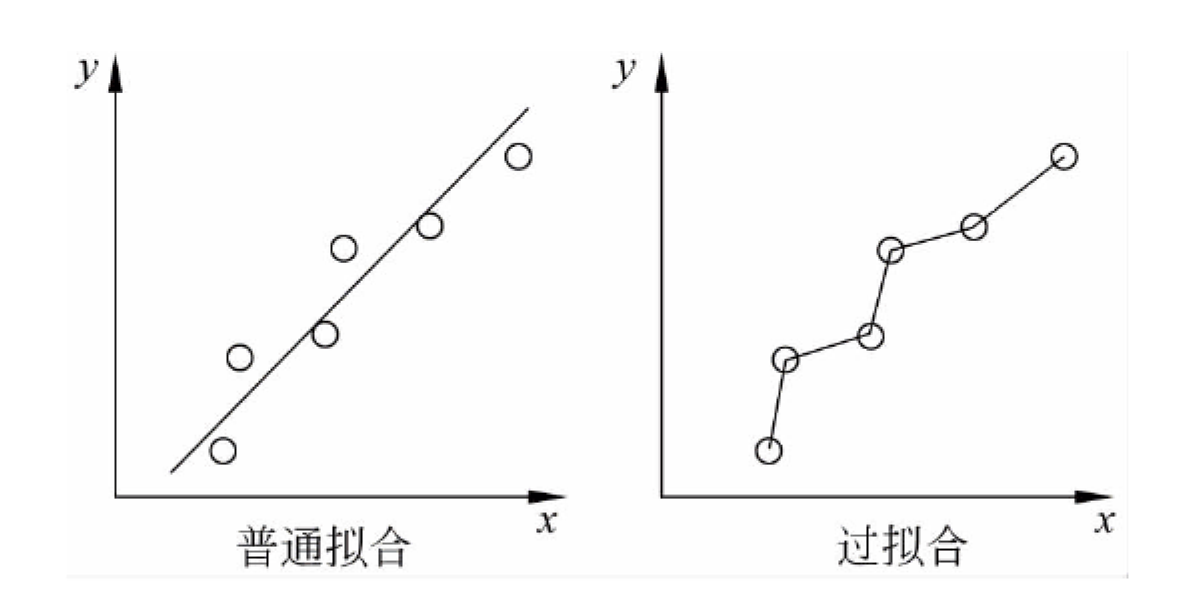


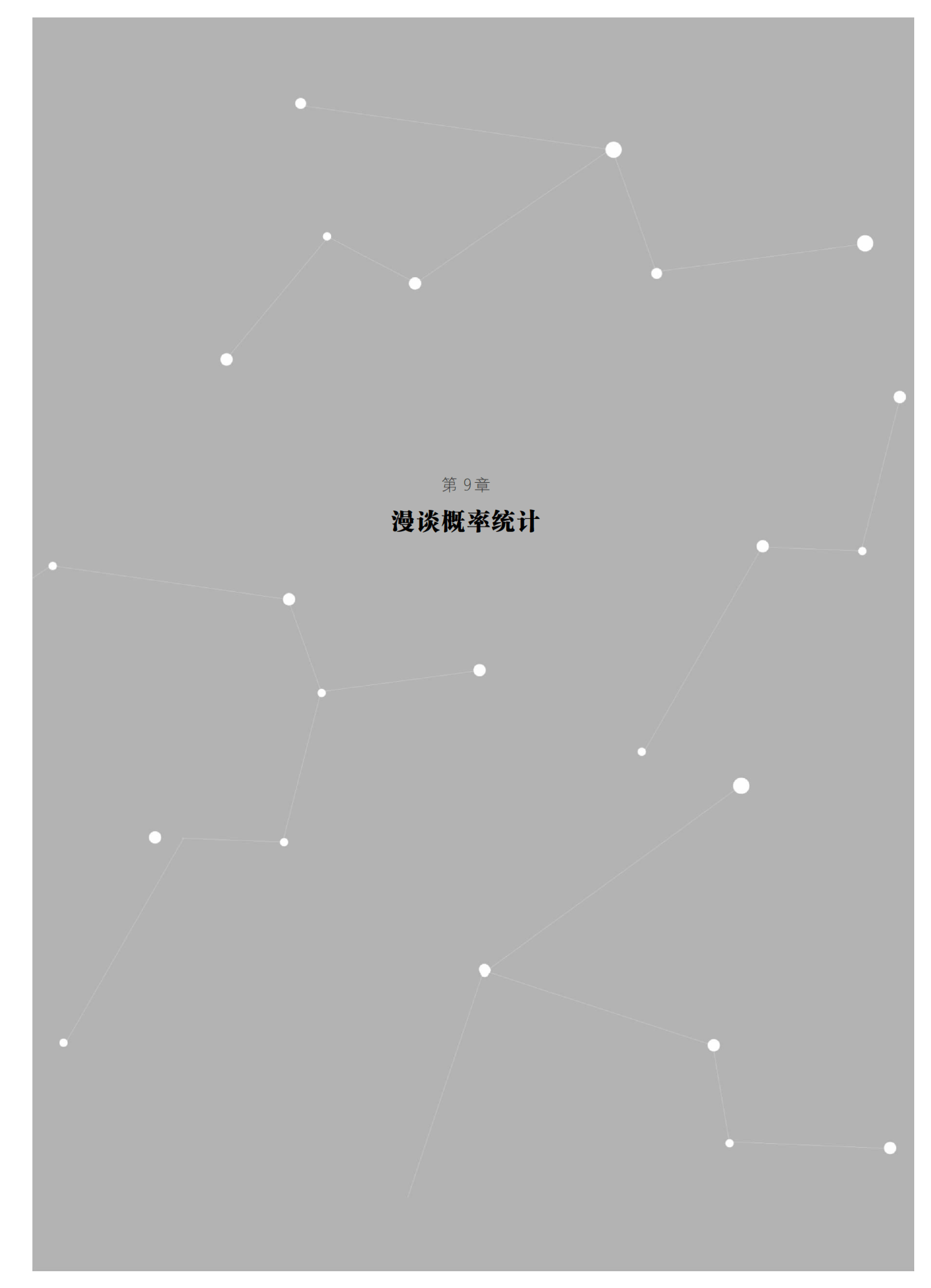
图 8-12　过拟合示意图

模型有效性

虽然谷歌的票房预测模型在2013年取得了成功，但这并不意味着该模型始终有效。谷歌的预测模型存在两个不稳定因素：一是前提条件不明确，任何预测方法都依赖前提条件，观众改变了对搜索引擎的使用习惯，或者电影预告片不再受宠，都会降低预测模型的准确率；二是相关关系不明确，如果线性回归中的参数 a 和 b 有明确的现实意义，模型的说服力会更强，也更容易辨识模型何时有效、何时无效，但在谷歌的票房预测模型中，票房和搜索量之间的线性相关关系是巧合还是必然，很难说清，这给模型带来了很大的不确定性，所谓的奇准预测可能只是昙花一现。

无论是线性回归，还是其他预测模型，我们都需要弄清楚模型的依赖条件和模型的现实意义，只有明确了这两点，才能明确预测模型何时有效、何时无效，从而避免模型的误用。

正所谓“理想丰满，现实骨感”，虽然利用数据预测未来让人着迷，却不易做到。以谷歌为代表的高科技公司正在带领我们揭开数据的神秘面纱，期待不久的将来，它们还会带给我们新的惊喜！



第 9 章

漫谈概率统计

导语：学了概率统计，应该懂得哪些常识？概率统计隐含了哪些元认知？常用的统计软件有哪几类？大数据究竟是什么？最后一章，我们一起来聊聊概率统计那些事儿。

9.1 正三观：概率统计常识

连岳是我非常喜欢的自由撰稿人，在不久前的“广东口腔科医生遭患者袭击身亡”事件之后，连岳撰文《真相好比凶杀现场》，对当下的医患关系做了一番评论，其中的一段是这样写的：

患者及医生普遍存在一种观念错误：

医生被美化为白衣天使，仿佛他们是另外一群特别神圣的人，不少医生也持这种自我认同。

……

医生应该改变的观念是：承认自己是自利人，不要再当天使，连比喻都不

接受。你要敢说："我努力学习、努力工作，就是为了多赚钱。"

患者的观念错误在于，许多人并不知道治病，也就是个概率事件，感冒不吃药，100%好，阿尔兹海默症，怎么吃药，100%好不了。很多危重疾病，似乎有些希望，但大家都尽力，可能也治不好。钱花光了，人又死了，一肚子的懊悔、心疼和怨气要找出气口，觉得医生神情可疑，自己听听传闻，用用搜索，受害情结越来越重——然后，只要1%的人失控，袭击医生的新闻，就不少了。

在一个环境里，双方不开心，冲突的可能性肯定增大。

这段精炼的评论里提到了一个医学常识：治病是个概率事件。有些病，比如感冒，治好的概率几乎是100%；而另一些病，比如阿尔兹海默症（俗称老年痴呆），以当下的医学水平，治好的概率几乎是0。清楚了这个概率常识，可以帮助我们正确看待绝症——即便医生和家属都全力以赴，治愈的可能性也微乎其微。

懂一点概率统计的常识，往小了说，可以让你变得更聪明，往大了说，可以矫正你的三观。下面，我们就来聊聊概率统计中几个不可不知的常识。

概率统计是"事后诸葛亮"

在抛硬币实验中，假如前9次都是正面朝上，第10次应当反面朝上了吧？

这是一个常见的认知错误——用概率统计结果做预测。这是因为人们普遍对大数定理心存误解。在抛硬币的问题中，提问者大约是这么想的：每一次抛掷，正反两面朝上的概率各为50%，如果前9次都是正面朝上，在随后的抛掷中出现反面的次数应该更多，否则就不符合50%概率的前提条件，因此第10次更可能是反面朝上。

这个想法犯了两个错误。其一，大数定理告诉我们，反复抛掷硬币多次，反面出现的次数占总次数的比例会越来越接近50%。注意，是"接近50%"，而不是"等于50%"。"接近"是相对的，2%比1%更接近50%，虽然它们都离50%远着呢；同时，"接近"是模糊的，是48%算接近，还是49%算接近，没人说得清。其二，大数定理是一个"描述性"的客观规律，所谓"描述性"指的是它只能事后描述抛掷结果，却无法决定任何一次抛掷的结果。在抛硬币实验中，每

一次抛掷都是独立事件，正反两面出现的概率永远各为 50%。你会不会猜中第 10 次抛掷的结果，只关乎运气。

条件改变概率

小镇昨夜发生了凶杀案，考虑到近 10 年来小镇只发生过 2 次凶杀案，应该很久不会再发生凶杀案了吧?

这是另一个常见的认知错误。虽然小镇平均 5 年才发生一次凶杀案，但是如果昨夜的案犯仍然在逃，小镇再次发生命案的概率将陡然提高，因为"案犯在逃"这个条件改变了凶杀案发生的概率。

条件概率是概率统计中最实用的概念，与之对应的贝叶斯定理则是最实用的计算公式。当我们需要计算某一个随机事件发生的概率时，类似事件的统计结果只能作为"先验概率"，尽可能多地掌握已知条件才能提高预测的准确率。

在"贝叶斯定理"一章中，我们曾提到，连环恐怖袭击不是巧合。仅从统计数据上看，"恐怖分子驾机撞世贸中心大楼"是不折不扣的小概率事件，然而在全球自杀式袭击数量飙升、基地组织越发猖獗的前提条件下，这一事件发生的概率则在悄然提升。当第一架被劫飞机撞向世贸中心大楼时，这一概率跃升至 38%，当第二架飞机再次撞向世贸中心大楼时，发生这一事件的概率飙升至 99%，几乎成为必然事件。事实上，第三架被劫飞机撞向了华盛顿五角大楼，第四架被劫飞机意图撞向白宫国会大厦，被机上乘客拼死阻止，最终坠毁。恐怖袭击总是连环发生，这不仅不是巧合，甚至是必然，正应了中国那句老话——祸不单行。

条件概率和贝叶斯定理提醒我们，不要盲目相信统计数据，前提条件会大大改变一个事件发生的概率。

均值不是唯一特征

每年国家统计局和各省市统计局都会发布"平均工资"的统计数据，媒体报道平均工资时常常使用"你拖后腿了吗?""你被平均了吗?"之类的标题，很

容易引起群众的热议。仅仅将自己的工资和平均工资作对比，就能知道自己有没有“拖后腿”吗？

表 9-1 是三组月薪调查数据，三组数据的平均值都是 10 000 元，于是我们告知被调查人员，平均月薪是 10 000 元，想想看，三组人员会有怎样的反应？第一组的大多数人会欣然接受这个结果，第二组是有人欢喜有人愁，第三组的大多数人会即刻加入“吐槽水军”，高呼自己“被平均了”。可见，均值相同并不意味着一切都相同，均值不是统计数据的唯一特征，标准差、最大（小）值、中位数等都是数据的特征，它们的作用是均值无法替代的。

表 9-1　三组月薪调查数据　　单位：元

人员编号	第一组	第二组	第三组
1	11 000	15 000	80 000
2	11 000	13 000	6 000
3	11 000	12 000	3 000
4	10 000	12 000	2 000
5	10 000	10 000	2 000
6	10 000	8 000	2 000
7	10 000	8 000	2 000
8	9 000	8 000	1 000
9	9 000	8 000	1 000
10	9 000	6 000	1 000

表 9-2 是三组数据的统计特征汇总表，从表 9-2 中可以看到，三组数据只有均值是相同的，其他统计特征各不相同，对比三组数据的统计特征可以得到新的认知。比如，第一组和第二组的标准差相比均值都较小，而第三组数据的标准差达到了24 640 元，是均值的近 2.5 倍，这说明第三组数据分布得极其分散，从最大值、最小值的对比也可以得到相似的推断。又如，第三组数据的中位数和四分位数都在 2 000 元、3 000 元附近徘徊，相比均值小很多，这说明有少数很大的数据将均值拉升到 10 000 元，反观第一组和第二组数据，就没有这种现象。

均值的确是数据的重要统计特征，但同时它只是一个统计特征，只有掌握了标准差、最大（小）值、中位数等多个统计特征，才能既全面又准确地解读出数据的内涵。

表 9-2　三组月薪数据的统计特征　　单位：元

统计特征	第一组	第二组	第三组
均值	10 000	10 000	10 000
标准差	816	2 867	24 640
最大值	11 000	15 000	80 000
最小值	9 000	6 000	1 000
中位数	10 000	9 000	2 000
四分位数	Q1：11 000 Q3：9 000	Q1：12 000 Q3：8 000	Q1：3 000 Q3：1 000

出场顺序无碍竞赛公平

当下的选秀节目五花八门，"中国好声音"寻找最美的声音，"最强大脑"寻找最聪明的大脑。无论哪个节目，参赛者都会按照抽签顺序依次出场，那么，出场顺序对参赛者的成绩有没有影响？第一个出场最不利，还是最后一个出场最不划算？我们一起来算一算。

三位选手 A、B、C 一同参加一个知识问答比赛，比赛规则是，选手从 20 张卡片中随机抽出一张，回答卡片上的 5 个问题，全部回答正确，就能赢得豪华双人游的机会。A 能答对 20 张卡片里的 9 张，那么，对 A 来说，第几个出场胜算最大？

如果 A 第一个出场，答对问题的概率很明显是 9/20。

如果 A 第二个出场，就需要分两种情况，前一个选手抽走了 9 张卡片中的一张，并且 A 答对问题的概率是(9/20)×(8/19)，前一个选手未抽走 9 张卡片中的任一张，并且 A 答对问题的概率是(11/20)×(9/19)，两个概率相加，A 答对问题的概率仍然是 9/20。

读者可以算一算 A 第三个出场时答对问题的概率，结果仍是 9/20。因此，仅从概率的角度来看，无论第几个出场，A 获胜的概率都一样，也就是说，出场顺序并不会妨碍比赛的公平。

9.2 元认知：概率统计之“道”

老子曰：“道可道，非常道。”意思是，道是可以被阐述的，但可以阐述的道不是真正的道。更接地气的说法是，道，只可意会，不可言传。老子所谓的“道”，是个抽象的指代，指的是“自然之道，万物之道”，这与当下的一个认知心理学概念颇为相似——元认知。“元”是本源之意，元认知指的是对认知的认知，比如学习如何学习、思考如何思考，它是方法背后的思想，技术背后的理念。每一门学科都可以提炼出元认知，这一节我们就来聊聊概率统计的元认知。

检验确保正确

小学一年级时，我们刚刚学习加减法，常常算错，老师会教我们做验算。如果是加法运算，就用结果减去加数，查看等不等于被加数，如果是减法运算，就用结果加上减数，检查是否等于被减数。

检验在数学中是必不可少的步骤，它帮助我们识别出错误的计算结果，提高正确率。假设检验是概率统计的常用检验方法，任何涉及统计量的计算，都需要对计算结果做假设检验，这在“假设检验”“线性回归”中都可以看到。只有经得起检验的结果才是正确可信的结果。

对比获得真知

佛说，要把一根绳子变短，只需找来一根更长的绳子。

在概率统计中，这句话蕴含的道理就是一个词——对比。正如上一节中平均工资的例子，仅仅知道平均工资的数值是远远不够的，要深入理解数据，就要做很多对比，不同城市的平均工资对比，同一城市不同行业的平均工资对比，平均工资与工资标准差对比，平均工资与工资中位数对比，等等。这些对比会加深我们的认识，帮助我们理解数据的内涵。

提防线性思维

问题 1：假定每一年都是 365 天，要使“至少两个人的生日为同一天”的概率达到 100％，至少需要多少人？

答：366 人。

问题 2：假定每一年都是 365 天，要使“至少两个人的生日为同一天”的概率达到 50％，至少需要多少人？

答：23 人。

我没写错答案，不是 183 人，是 23 人。计算过程如下所述。

两个人时，要使他们的生日不同，只需让第二个人的生日避开第一个人，所以概率是 364/365，两人生日相同的概率是 $1-364/365=0.003$。

三个人时，要使他们的生日不同，需要第二个人的生日避开第一个人，同时第三个人的生日避开前两个人，所以概率是：

$$1-(364/365)\times(363/365)=0.01$$

按照这个方式便可以计算 n 个人中至少两人同一天生日的概率是：

$$1-(364/365)\times(363/365)\times(362/365)\times\cdots\times(366-n)/365$$

当 $n=23$ 时，这个概率便超过了 50，因此第二个问题的答案是 23。

之所以很多人认为是 183 人，是因为他们把第二个问题想成了“至少一个人与你的生日相同，至少需要多少人”。两个问题的不同点在于，“与你生日相同”是线性的，“至少两人生日相同”不是线性的，是网状的。试想 A、B、C 三个人的情况，B 或 C 与 A 同一天生日满足问题中的条件，同时 B 和 C 同一天生日也满足。4 个人、5 个人的情况将更复杂，每个人都可能与其他人生日相同，这将构成一个庞大的概率网络，必定不能用线性思维去解释。

回到第一个问题，为什么答案是 366 人？因为问题中的说法是“达到 100％”，而不是“接近 100％”。利用上面的公式可以计算出，当 $n=50$ 时，至少两人生日相同的概率就会达到 97％，十分接近 100％，人数的进一步增加只会把这个概率缓慢地推向 100％。

这是概率统计中经典的生日谜题，它提醒我们，简单的线性思维很可能出现错误，在解答问题前，要给问题定性，只有线性的问题才能用线性思维求解。

总是反过来想

投资大师查理·芒格曾在演讲中提到一个乡下人的故事，这个乡下人说："要是我知道我会死在哪里就好了，这样我就永远都不会去那个地方。"看似调侃的一句话包含了查理·芒格最重要的思维方式，他称为"总是反过来想"。

在概率统计中，我们称为"反证法"。当你要证明某个参数等于某个数值时，最好的办法就是反证法，首先假设等于关系成立，再由此得到推论，如果推论与已知条件存在矛盾，说明假设是错的，即等于关系不成立，反之则成立。假设检验正是沿用了反证法的思路，唯一不同的是，假设检验是以显著性水平的形式作出判断，但这并不影响反证法本身。

"如果我不能比全世界最聪明、最有能力、最有资格反驳这个观点的人更能够否定这个观点，我就不配拥有这个观点。"这是查理·芒格的另一句名言，可谓逆向思维的最高境界，在此送给读者，与君共勉。

模糊的正确胜过精确的错误

沃伦·巴菲特："我宁要模糊的正确，也不要精确的错误。"

经历过 2008 年金融危机的人，都会明白巴菲特这句话的含义。就在中国股市如火如荼之时，巴菲特却在以 13 港元的价格陆续减持中石油 H 股，后来中石油在回归 A 股的利好刺激下冲高至 20 港元，巴菲特因此错过了将近 50%的收益。大浪淘沙，只有时间能说明一切。2008 年年底，没有人再会嘲笑巴菲特损失的"区区"50%收益，相比于 50%的收益，50%的损失对投资者的伤害要大得多。"模糊的正确"，是给股票的内在价值划定一个区间，这胜过一个貌似精确实则错误的数字，这就是巴菲特的哲学。

在概率统计中，也会有很多模糊的说法，比如，"二八法则"指的是指数分布的特征，未必要精确符合 20%的人掌握 80%的财富这个比例，又如，为了节约计算成本，我们常常使用泊松分布代替二项分布。有时，我们放弃了"精确"，却可以得到"正确""安全""快捷"甚至更多。

9.3　兵器谱：统计软件大盘点

工欲善其事，必先利其器。行走江湖，行侠仗义，一件称手的兵器是必不可少的。在当下的信息时代，统计软件就是统计分析人员必不可少的兵器。下面我们就来列举一些统计软件的“兵器谱”。

统计软件可以分为通用软件、商用软件和开源软件三类(如图 9-1 所示)。



图 9-1　常用统计软件

通用软件

通用软件毫无疑问指的是 Excel。Excel 是微软办公套件中的一个组件，适用于 Windows 平台，可以用于数据处理、统计分析和图表绘制，在管理、财务、金融等诸多领域被广泛使用，是众多职场人士的必备软件。在统计分析方面，Excel 可以计算数据的统计特征(均值、方差等)，绘制各类统计图表(散点图、柱状图、饼图等)，还可以进行初级统计分析(方差分析、线性回归等)。便捷的操作是 Excel 的一大优势，但是如果你要处理成千上万行的数据，这一优势会瞬间消失：一来庞大的数据会占用大量内存，导致软件卡顿；二来你不得不花费大量时间练习使用甚至自定义大量的快捷键，以应付屏幕上无法显示全部数据的尴尬局面。所以说，Excel 适合对少量数据做简单的统计分析。

Numbers 软件是 MAC 平台上的数据处理软件。它既可以用于计算数据的统计特征，还可以绘制各类 2D 和 3D 的图表，但是不具备统计分析功能。与 Excel 相似的是，Numbers 也不适合处理大量数据。Numbers 和 Excel 是单向兼容的，Numbers 文件可以保存为 Excel 文件，反向则不支持。

商用软件

商用软件指的是 SPSS、SAS 和 BMDP，三者并称世界三大统计软件包，是为统计分析人员打造的专业工具，均为付费软件。

统计产品与服务解决方案（Statistical Product and Service Solutions，SPSS）。1968 年，美国斯坦福大学的三位研究生开发完成了 SPSS——世界上最早的统计分析软件，同时成立了 SPSS 公司。2009 年 7 月，IBM 公司收购了 SPSS 公司，现在 SPSS 软件属于 IBM 公司的产品。SPSS 的界面风格与 Excel 类似，提供了非常丰富的统计分析模型，包括时间序列分析、逻辑回归、聚类分析等高阶分析工具，并可以输出各种精美的图表，主要运行于 Windows 平台。SPSS 广泛应用于社会科学、自然科学的科学研究和工程实践中。

统计分析系统（Statistical Analysis System，SAS）。SAS 是由美国北卡罗莱纳州立大学于 1966 年研发出的专业统计软件，1976 年 SAS 软件研究所成立，负责 SAS 软件的维护、开发、销售和培训工作。SAS 是一个模块化、集成化的大型应用软件系统，包含数据访问、数据储存及管理、应用开发等十几个模块，可以完成数据访问、数据管理、数据呈现和数据分析四类任务。SAS 主要应用于政府、管理、科研、金融等领域，我国的国家信息中心、国家统计局、中科院等单位都是 SAS 的用户。

生物医药数据处理（Bio Medical Data Processing，BMDP）。BMDP 由美国加州大学洛杉矶分校于 1961 年研发而成，是由一个名为 BIMED 的生物医学软件修改而来。1968 年 BMDP 公司成立并发行 BMDP 软件，当时 BMDP 是国际知名的综合专业统计分析软件，有很多独具特色的分析方法。可惜 BMDP 公司发展不顺，最终被 SPSS 公司收购，BMDP 也失去了昔日的光辉，在与 SAS 的竞争中处于劣势。

开源软件

在开源软件领域,用于统计分析的有R和Python两个编程语言。

R语言是用于统计分析和绘图的专用编程语言,是一个自由、免费、源代码开放的编程语言,其源代码托管于github。R语言诞生于1980年左右,是S语言的一个分支,S语言是美国AT&T公司贝尔实验室开发的统计分析语言,后来新西兰奥克兰大学的开发团队研发出了首个R语言运行系统。除了免费开源,R语言还是一个跨平台的语言,可以用于UNIX、Windows和MacOS三类主流操作系统。借助开源社区的不断发展,R语言正在收获越来越多的功能扩展包,在金融分析、科学研究和人工智能等领域的应用也越来越广泛。本书的大多数统计图线都是用R语言绘制的。

Python是一个面向对象的、解释型的编程语言,诞生于1991年,目前广泛应用于系统管理和Web编程。严格地讲,Python并不是用于统计分析的编程语言,但Python拥有异常强大和丰富的函数库,借助Numpy、Scipy、Matplotlib等函数库,可以实现大多数统计分析和绘图功能。Python与C、C++和Java等常用编程语言可以完美结合,因此,Python是程序员们进行统计分析的首选工具。

9.4　大数据：创新与挑战

仿佛一夜之间,“大数据”成了家喻户晓的常用词,不论新兴行业还是传统行业,都准备“拥抱大数据”,都想从大数据中发现宝藏。可是,大数据究竟是什么?是新瓶装旧酒,还是技术革命?本书的最后一节,我们来探一探大数据的底。

大数据的概念可以追溯到2001年,世界知名咨询公司Gartner发布的一份咨询报告首次提出“Big Data”,并提出了“3V”模型,意思是大数据在数量(Volume)、速度(Velocity)和种类(Variety)三个维度上都很“大”。但是受限于当时的软件技术,大数据只能停留在概念层面。进入21世纪的第二个十

年，随着并行计算和数据分析技术的兴起，大数据终于迎来了大爆发时刻。2012 年，畅销书《大数据时代》令“大数据”一词迅速普及，各行各业都对大数据技术跃跃欲试。大数据技术在互联网、娱乐等行业率先得到应用，很多应用成果令人耳目一新，比如美剧《纸牌屋》的策划、巴西世界杯的预测。

大数据含义丰富，难以定义，目前比较权威的定义是 Gartner 给出的：“大数据是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产”。这个定义包含了大数据的三个典型特征：

新形态——大数据是海量、高增长率和多样化的信息资产；

新模式——大数据需要新处理模式来处理；

新能力——大数据具有更强的决策力、洞察力和流程优化能力。

新形态

大数据最鲜明的特征自然是“大”，即海量的数据。截至 2016 年年初，全球网民数量达到 34 亿，移动用户更是达到 37.9 亿，超过全球总人口的一半；中国的社交网络工具——微信，在 2015 年创下了月活跃用户破 6.5 亿的记录；2015 年 11 月 11 日，阿里巴巴网上销售平台全天销售额达到创纪录的 912 亿元。庞大的互联网用户群体不停地生产着数据，这就是海量数据的源头。未来，随着物联网的普及，全球所有设备都会为互联网贡献数据，那时全球互联网的数据量将超出你我的想象。

海量的数据要靠高增长率才能实现。看小说、看视频，微信聊天、淘宝购物，每一个网民都在不停地为互联网贡献流量。在中国第二届大数据产业峰会上，美国高通公司全球总裁德里克·阿伯利在演讲中提到：“现在的数据是呈指数级发展的，过去两年产生了全球 90％的数据量。”指数级的高增长率正是大数据的又一鲜明特征。

在大数据技术兴起之前，人们习惯于把数据存储在关系型数据库中。关系型数据库就像一堆大型 Excel 表格，每个表格都有很多列，每一列代表数据的一种属性，数据按照对应的属性存储起来，可以互相关联，便于查找。比如，公安局中存储的公民信息，会列出姓名、性别、身份证号、家庭住址、联系电话等属性，然后把每个人的信息录入数据库中保存起来。这种属性划分明确的

数据称为结构化数据。随着互联网的普及，网络上的信息门类越来越丰富，E-mail、新闻报道、聊天记录、自拍图片、自拍视频等，网友们在互联网上自由分享着这些零散的、随性而成的信息。很显然，这些信息并不适合存储在关系型数据库中，因为这些数据是非结构化的。非结构化数据和结构化数据的混合共存是大数据的又一特征——多样化。

新模式

传统数据库技术无法高效处理海量、高增长率、多样化的大数据，革命性的新处理模式应运而生。2003 年，谷歌发表了题为 *The Google File System* 的论文，向全世界介绍了它们设计实现的分布式文件系统 GFS（Google File System），在 GFS 的基础上，谷歌提出了并行处理架构"MapReduce"和分布式数据存储系统 Bigtable，这三个软件是大数据"新处理模式"的典型代表。受到谷歌的启发和激励，开源软件基金会 Apache 开发出了 Hadoop 系统，它包括分布式文件系统 HDFS（Hadoop Distributed File System，HDFS）和 Map/Reduce 并行处理两部分。Hadoop 引领了大数据处理模式的革命浪潮，Hive、HBase、Spark、Storm 等开源软件相继出现，形成百家争鸣的局面。

从原理上讲，GFS 和 HDFS 很相似，二者都是分布式的，都可以部署在廉价硬件集群上，都具有良好的容错特性。MapReduce 则将数据处理分为"映射（map）"和"归约（reduce）"两个独立的步骤，实现了海量数据的并行处理。Spark 弥补了 Hadoop 高延时的缺陷，实现了高速的并行数据处理。Storm 是推特公司使用的"流式处理"系统，适用于处理不断产生的实时消息，即流式数据。2015 年，推特公司用新方案 Heron 替代了 Storm，大大提高了吞吐量并减少了硬件开销。

上述软件系统是大数据处理新模式的典型代表，随着大数据处理需求的增加，必定还会有更多的新软件、新系统出现。

新能力

大数据是创新，更是革命，海量的数据不仅可以用作统计分析，还可以用

作产生“智慧”。

凯文·凯利在《失控》中曾提到，当高度互联的低级群体的数量大到一定程度时，群体特征便会涌现出来，该特征是群体中的任何个体都不具备的。比如，大量水滴汇集成河水、海水，便会产生让水滴“感到陌生”的新特征——旋涡和波浪。大量机器聚集起来能否涌现出智慧？这个曾经的哲学问题被数据科学家解决了——机器不仅会拥有智慧，而且会越来越聪明，因为人类赋予了机器学习的能力。

十几年前，沃尔玛超市从销售数据中发现“啤酒和尿布”的关联关系，令世人震惊。如今我们回头去看，这只是机器学习中十分简单的关联算法。机器学习，即让计算机具有学习能力。近几年来，伴随着数据量的高速增长，可供计算机学习的素材越来越多，机器学习的各种算法也迅速发展和普及。邮件服务器可以自动识别垃圾邮件，亚马逊网站自动向你推荐“你可能喜欢的”商品，量化投资基金通过高频交易赚取利润，公安局利用监控录像识别嫌疑人身份，贝叶斯分类器、逻辑回归、Apiriori 关联等机器学习算法得到越来越多的应用，大数据时代就是机器学习的时代。

2016 年 3 月 15 日，谷歌围棋人工智能程序 AlphaGo 以 4∶1 的总比分战胜了韩国棋手李世石，令世人哗然。AlphaGo 是如何炼成的？答案是深度学习。深度学习是机器获得智慧的另一种方法，它模拟人脑神经网络的学习模式，实现由简单到复杂的学习过程。简言之，深度学习将使机器拥有创造力甚至想象力！

在机器学习和深度学习的辅助下，大数据正在涌现智慧，这正是大数据具备的新能力——更强的决策力、洞察力和流程优化能力。

新挑战

新挑战是我为大数据加入的第四个特征。

大数据带来了创新甚至革命，也同样面临严峻的挑战。大数据常常挖掘数据间的相关性，可是相关性有没有意义，相关性是不是可靠，都应当受到质疑。比如，大数据分析会发现从 2006—2011 年，美国谋杀案比例与 IE 浏览器的市场份额有很高的相关性，都呈急速下降趋势，但是这样的相关性有什么意

义，很难说得清。又如，谷歌流感趋势预测系统在刚刚推出时能够准确预测流感趋势，可是 4 年后就出现了巨大的错误，其预测的就诊数据比实际数据高出两倍之多，而且这种失准持续了很久也无法得到改善。

大数据面临的另一个挑战是噪声。数据量的增加会让分析结果更精确，但精确不等于正确，海量的数据会引入海量的噪声，这些噪声会淹没有效信号。就在“9・11”恐怖袭击发生前的几个月，美国联邦调查局探员肯・威廉姆斯发现，近几年亚利桑那州的多家飞行学院涌入了很多学员，他对这些学员进行了背景调查，发现他们大多与基地组织有关联，于是他给联邦调查局提交了一份报告，提到了基地组织可能正在将一些学生送到美国的各所飞行学院去学习，这些学员一旦进入民航系统，可能会借机发动恐怖袭击。这份报告被标注为“普通”和“只是一种猜测，不是很重要”，最终湮没在联邦调查局堆积如山的报告中，“9・11”事件发生后，人们称为“凤凰城备忘录”。大数据分析同样可能出现“凤凰城备忘录”式的悲剧，有价值的信号湮没在巨大无比的噪声中。

大数据时代，既是创新，也面临挑战。人类从未如此高度地互联，人类也从未如此高速地生产数据，属于大数据的时代正在缓缓地拉开大幕，让我们拭目以待吧！

参 考 文 献

[1] 盛骤,谢式千,潘承毅.概率论与数理统计[M].北京:高等教育出版社,2008.

[2] [美]Hadley Wickham. ggplot2. 数据分析与图形艺术[M]. 统计之都,译. 西安:西安交通大学出版社,2013.

[3] [美]Ronald J. G. 让你爱上数学的 50 个游戏[M]. 庄静,译. 北京:机械工业出版社,2015.

[4] [瑞典]Peter Olofsson. 生活中的概率趣事[M]. 赵莹,译. 北京:机械工业出版社,2014.

[5] [美]Nate Silver. 信号与噪声[M]. 胡晓姣,等,译. 北京:中信出版社,2013.

[6] [美]Dawn Griffiths. 深入浅出统计学[M]. 李芳,译. 北京:电子工业出版社,2012.

[7] [美]Michael Milton. 深入浅出数据分析[M]. 李芳,译. 北京:电子工业出版社,2012.

[8] [美]Drew Conway,John Myles Wbite. 机器学习实用案例解析[M]. 陈开江,等,译. 北京:机械工业出版社,2013.

[9] Ginsberg J,Mohebbi M H, Patel R S, et al.. Detecting influenza epidemics using search engine query data[J]. *Nature*, 2009, 457(7232): 1012-1014.

[10] Reggie P, Andrea C. Quantifying Movie Magic with Google Search[J]. Google Whitepaper-Industry Perspectives+ User Insights.

[11] [美]William Poundstone. 推理的迷宫[M]. 李大强,译.北京:中信出版社,2015.

[12] [美]Ambrose Bierce. 鹰溪桥上[M]. 程闰闰,译. 重庆:重庆大学出版社,2013.

[13] [美]Peter D. Kaufman. 穷查理宝典[M]. 李继宏,译. 上海:世纪出版集团, 2012.

[14] [美]Kevin Kelly. 失控[M]. 张行舟,等,译. 北京:电子工业出版社,2016.

[15] [美]Kevin Kelly. 必然[M]. 周峰,等,译. 北京:电子工业出版社,2016.